Incorporating high-dimensional exposure modelling into studies of air pollution and health

Liu, Yi

Award date:
2015

Awarding institution:
University of Bath

Incorporating high–dimensional exposure modelling into studies of air pollution and health

submitted by

Yi Liu

for the degree of Doctor of Philosophy

of the

University of Bath

Department of Mathematical Sciences

December 2014

Summary

Air pollution is an important determinant of health. There is convincing, and growing, evidence linking the risk of disease, and premature death, with exposure to various pollutants including fine particulate matter and ozone. Knowledge about the health and environmental risks and their trends is important stimulus for developing environmental and public health policy. In order to perform studies into the risks of environmental hazards on human health study there is a requirement for accurate estimates of exposures that might be experienced by the populations at risk. In this thesis we develop spatio-temporal models within a Bayesian framework to obtain accurate estimates of such exposures. These models are set within a hierarchical framework in a Bayesian setting with different levels describing dependencies over space and time. Considering the complexity of hierarchical models and the large amounts of data that can arise from environmental networks mean that inference using Markov Chain Monte Carlo (MCMC) may be computational challenging in this setting. We use both MCMC and Integrated Nested Laplace Approximations (INLA) to implement spatio-temporal exposure models when dealing with high–dimensional data. We also propose an approach for utilising the results from exposure models in health models which allows them to enhance studies of the health effects of air pollution. Moreover, we investigate the possible effects of preferential sampling, where monitoring sites in environmental networks are preferentially located by the designers in order to assess whether guideline and policies are being adhered to. This means the data arising from such networks may not accurately characterise the spatial-temporal field they intend to monitor and as such will not provide accurate estimates of the exposures that are potentially experienced by populations. This has the potential to introduce bias into estimates of risk associated with exposure to air pollution and subsequent health impact analyses. Throughout the thesis, the methods developed are assessed using simulation studies and applied to real–life case studies assessing the effects of particulate matter on health in Greater London and throughout the UK.

Acknowledgements

Firstly, I would like to thank my supervisor Dr. Gavin Shaddick for his patient guidance and friendship throughout my PhD study. I would also like to thank all the friends I have made during the last three years. My deepest thanks goes to my parents, who have always been there for me, their love and support have helped me through the hard times.

Declaration

The work presented in Chapter nine forms the basis of the work in the paper '*Incorporating high-dimensional exposure modelling into studies of air pollution and health*' with Dr. Gavin Shaddick and Prof. Jim Zidek which has been submitted to the Statistics in Biosciences. This work, together with some of the material in chapter 10, will also be presented at the 24th annual meeting of the International Environmetrics Society (TIES) in Ghangzhou in 2014, with the titles *Incorporating large–scale exposure modelling into studies of air pollution and health*, and *Mitigating the effects of preferentially selected monitoring sites for inference and policy*, both jointly authored with Dr. Gavin Shaddick and Prof. Jim Zidek.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Numerous research studies across the world have shown that exposure to poor air quality impacts on people's health. Some effects are nearly immediate, known as acute effects, while some happen over a longer term, known as chronic effects. Considering the very large number of people who live in urban areas and therefore may be be exposed to substantial amounts of air pollution, this is an important and in places urgent health issue. The World Health Organization estimates that 7 million deaths each year may be directly attributable to air pollution (WHO, 2011).

Poor air quality in Britain, especially in major cities such as London has been a problem for hundreds of years. Historical documents indicate there was a law enacted by King Edward I to control coal burning in 1306. After the industrial revolution in 18^{th} century, coal smoke and its associated problems were considered to be a serious issue in London up until late 20th century. The most infamous incident caused by coal burning was the London smog in 1952, often referred to as the 'Great Smog', in which levels of black smoke, a measure of particulate matter, exceeded 4,500 μgm^{-3}. During this episode it is estimated that 4,000 people died prematurely and 100,000 more were made ill due to effects on the respiratory tract (Logan, 1953). Nowadays, the sources of pollutants within the air have changed considerably, due largely to the marked decline in the use of coal for industrial processes and domestic heating over the past 40 years. At the same time, the rapid expansion in the number of motor vehicles has produced considerable amounts of nitrogen dioxide and small particles. A wide range of pollutants have been implicated in adverse effects on human health, but particular attention has tended to focus on particulate matter, measured in various ways, e.g. PM_{10}, $PM_{2.5}$, total suspended particulate and black smoke.

Particulate matter consists of tiny bits of solids or liquids suspended in the air, the majority

of which are too small to see. They are a complex group of pollutants varying in size, shape, composition and origin, and vary from place to place and across time. Others kinds of particles include material from building and industry, sea salt, pollens and soil particles. Among these different types, size is the main determinant of which part of body the particles harm when breathed in. Larger particles generally rest within the nose and throat, but particulate matter smaller than 10 micrometers, referred to as PM_{10}, can settle in the lungs (Elliott et al., 2007). Small particles of particular matter (PM) in particular are now considered to be a major source of risk to health. Much of the evidence for this has come from epidemiological studies.

Epidemiological studies require accurate measurements of both health outcomes, potential confounders and estimates of exposures that might drive associations with health (Finazzi et al., 2013). All of these data may be measured with varying degrees of error. Considering measurements of air pollution, there is a true underlying pollution surface which will form the basis of the exposures experienced by the population at risk. However this surface is not directly observable and instead measurements are taken at locations over space and time. Differences between these exposure measurements and the unknown underlying field are often referred to as measurement error. Here the term measurement error is taken to refer to any difference from the underlying true values (of pollution) and what is measured. Traditionally measurement error has been based around the idea of repeated measurements of a value, for example measuring blood pressure, in which the repeated measurements will contain a component of error; often assumed to be random. In modelling exposures and in spatial epidemiology, error will possibly comprise a number of factors including monitor calibration error and random variation but also variations in the underlying pollution field over time and space which are not acknowledged in the analyses. This may arise for example when modelling assumptions are too simplistic for the complex surface of the pollution field. Another issue that is often contained under the umbrella term of 'measurement error' is the misalignment of locations or times of exposure measurements and health outcomes. This arises because exposure and health data are often drawn from independent sources and not as the result of a carefully designed study. Hence a straightforward comparison is not possible without a model to align these elements in the spatial and temporal domains (Gryparis et al., 2009; Peng and Bell, 2010). Where a health effects analysis uses predictions from an exposure model as substitutes for actual measures of exposures, as with regular measurement error, there is the possibility of bias in the estimation of risks. An additional issue termed the 'change of support' problem by Gelfand et al. (2001) occurs when the exposure and health outcome data are recorded at different levels of aggregation, for example health counts for administrative areas and exposures from monitoring sites at point locations within, or outside, those areas. The studies are ecological in nature, being based on spatially aggregated health and exposure data modelled at the same resolution. As

such, there is the potential for ecological bias; assuming that associations observed at the level of the area hold for the individuals within the area. For a comprehensive review of the problems of ecological bias and possible approaches for corrections, see Wakefield and Salway (2001), Wakefield (2003) and Wakefield and Shaddick (2006).

1.1 Estimating short–term effects

The relationship between exposure to air pollution and mortality or morbidity has been an active research topic for a number of years. Much of the evidence about the effects of air pollution comes from studies of acute health effects, that is, from associations between short-term changes in air pollution and subsequent changes in mortality or morbidity. The majority of studies (Samet et al., 2000; Katsouyanni et al., 2001) regress population based mortality counts for an area, e.g. a city, against ambient pollution levels. The pollution levels are often collected at a number of fixed locations, with the majority of studies using a 'standard' measure of daily pollution exposure; the daily average from measurements from all available monitoring sites. A number of studies have used a spatio-temporal pollution model in this setting, largely due to the health data being available at a lower geographical temporal resolution than the exposures data, and because pollution concentrations were not available in each spatial unit. For examples, Zidek et al. (1998), Zhu et al. (2003), Fuentes et al. (2006), Lee and Shaddick (2010), Szpiro et al. (2011) and Chang et al. (2011) used spatio-temporal models to produce less biased estimates of both exposure and the association with mortality. Also simple methods for handling missing values are commonly used including simply discarding them from the analysis or replacing them by a specific single value, for example the overall mean. By discarding missing values, we may lose useful information and may introduce bias. When replacing missing values by a single value, for example a sample mean of observations or the posterior mean from an exposure model the intrinsic variability associated with the summary value may be ignored. In Chapters 4 we assess the estimation of short–term effect of air pollution. We start by considering the standard model, a Poisson log-linear model with exposures comprising daily averages of measurements from available monitoring sites. We assess how well it performs, and assess potential biases, when there is measurement error and spatial variation in the data. This is done using simulation studies. We assess the potential benefits of using more complex approaches and models in Chapter 6, including a measurement error model and a spatial–temporal model, both set within a Bayesian framework with inference using Markov Chain Monte Carlo (MCMC). Their ability to accurately estimate risks is assessed using simulation studies. These models are also applied to a case study of data from London in Chapter 8.

1.2 Estimating longer–term effects

Compared to short–term effects, there has been comparatively much less research into the chronic effects of air pollution, i.e. the association between health outcomes and long–term exposures to air pollution. This has been due to a number of reasons, including lack of availability of suitable data and problems of confounding. It is not really known whether short–term effects can be extrapolated to longer term effects. Whilst some of the acute effects might just be bringing forward health events which were likely to occur within a short time anyway known as 'mortality displacement', long–term exposures may be fundamental in causing disease, e.g. by sensitising people in early life to respiratory allergens (Elliott et al., 2007).

Two early studies were particularly important in indicating the long–term effects of air pollution. The Harvard Six Cities Study (Dockery et al., 1993), followed a cohort of over 8000 adults in six cities in the U.S., from the mid 1970s to the late 1980s. Exposure data were obtained for each city using a single pollution monitor in the center of the city which was then linked with health and confounder data obtained from questionnaires and interviews. Strong associations were found between mortality rates in the six cities and concentrations of particular matter and sulphate particles. Twenty-six percent higher all–cause mortality was found between the cities with the higher levels of pollution component to the lowest. Similar risks were seen in the American Cancer Society (ACS) study, which contained 151 cities (Pope et al., 1995).

In Chapter nine, we investigate models for assessing the longer term effects of air pollution; looking at the associations between health and exposures possibly over several years. In many epidemiological studies, where there are missing values in exposure information very simple methods are used including simply omitting them, which may result in omitting entire data records, to replacing them by a single value, for example the overall mean. Important information may be lost if records are discarded and important features of the data may be ignored by replacing missing values by a single value. In addition, the intrinsic variability in using a summary value is commonly ignored. A more advanced way of dealing with missing data is to use a model for exposures and then to 'fill in' missing values with predictions from the model. However, when dealing with the large amounts of data that may arise from considering exposure measurements over a long period of time there may be computational issues when attempting to perform inference using MCMC. Therefore, we investigate the use of using approximations, namely Integrated Nested Laplace Approximations (INLA) in this instance. We propose an approach for integrating the results of a space–time exposure model into a health analysis, which means using the output of the exposure model (predictions) as input for the

health model. This is illustrated using a case–study of black smoke exposures and respirtatory health in the UK over an extended period of several decades.

1.3 Preferential sampling

Early air pollution control legislations were focused on setting restrictions on the use of smoke-producing fuels (Stern, 1973) and in 1961 the world's first national air pollution monitoring network was established in the UK. It was called the National Survey and monitored black smoke, a measure of PM and sulphur dioxide at around 1000 sites (Clifton, 1964). Since then many countries have established monitoring networks. During the early parts of the twentieth century, the main concern was soot (or black smoke) and sulphur dioxide from industry and domestic fires and at that point networks were largely designed in order to measure these pollutants with many monitors being located in industrial areas where concentrations were likely to be high. Now, concern may be focused on PM due to road traffic and this will drive the locations of many monitoring sites, ie. beside roads.

Following legislation at both national and international levels and air quality guidelines (AQGs) from the World Health Organisation (WHO), monitoring air pollution has dramatically increased. The AQGs aim to offer guidance to reduce the health impacts of air pollution. However, the information that is available to support air pollution policy has three main things wrong with it; (i) monitoring is expensive and so monitoring networks do not cover every area (ii) concentrations may vary greatly over small distances, especially in urban areas and (iii) networks are often designed to monitor compliance with standards and therefore may mostly be in areas with high pollution and so may not give accurate representation of true levels of pollution that might be experienced by populations.

It is very important that the information coming from networks is accurate and reflects the levels of exposures that may be experienced by the populations at risk. This might be a problem if monitors are placed in locations where pollution might be expected to high; known as *preferential sampling*. In the context of air pollution and health in epidemiological analyses, Guttorp and Sampson (2010) state that the choice of locations for air pollution monitoring sites may be because of a number of reasons, including measuring: (i) background levels outside of urban areas; (ii) levels in residential areas and (iii) levels near pollutant sources. Geostatistical methods which assume sampling is non-preferential are often used despite preferential sampling (Diggle et al., 2010). Ignoring preferential sampling may lead to incorrect inferences and biased estimates of pollution concentrations and thus any subsequent estimation of health risks.

In chapter ten, we assess the potential effects of preferential sampling on the estimation of health risks associated with air pollution. Preferential sampling may lead to inaccurate estimates of exposures and this in turn has the potential to introduce bias into estimates of risk. Following on from the approach for incorporating exposure models in health studies presented in Chapter 9 we consider the possible effects of preferential sampling on risk estimates. We assess these possible effects using a simulation study under which estimates of health effects are compared when using random and preferential sampled sets of exposures. We also present a case study based on a long–term air pollution monitoring network in the UK which has previously been shown to have been subject to preferential sampling over time (Shaddick and Zidek, 2014). We propose a method to adjust for preferential sampling by using predictions from an exposure model based on non-preferentially sampled data in place of preferentially sampled data.

1.4 Computation

Throughout this thesis we compare the performance of models using simulation studies and by applying them to real data. In spatial–temporal applications, where data is available from a large number of monitoring sites over long periods of time, the dimension of the data that needs to be handled may become very large. Although in theory Markov chain Monte Carlo method can be applied to all of the Bayesian hierarchical models considered in this thesis, it may come with a very heavy computational burden. As an alternative to MCMC, Integrated Nested Laplace Approximation (INLA) has recently been introduced to perform fast Bayesian inference. This approach obtains results by using numerical approximation of the marginal posterior densities of variables of interest and hyperparameters instead of simulation, and as such it can be very computationally efficient.

Chapter 2

Exposure Modelling

If the measurements of pollution are mis-estimated or inappropriate, then estimates of the relationship with mortality may be biased. It is important to understand characteristics of the pollution surface, and in particular levels of measurement error and spatial variation. In this Chapter, we introduce the exposure modelling framework adopted in this thesis. The first section presents details of spatial processes and the following section introduces temporal processes. Section 2.3 outlines the basic idea of spatio-temporal modelling. Then some examples of applications of spatio-temporal modelling in environmental modelling are given. Lastly, Section 2.5 introduces measurement error commonly seen in exposure models.

2.1 Spatial processes

A spatial random field is a stochastic process over a region $Z = \{Z_s \subset \mathcal{R}^d\}$ for $s = 1, \ldots, N_s$. This underlying process is not directly measurable, but realisations of it can be obtained by taking measurements, possibly with error, at a set of known locations. One way of expressing the random field is as a combination of an overall trend together with a spatial effect, for example

$$\begin{aligned} Y_s &= Z_s + \epsilon_s \quad for\ s = 1, \ldots, N_s \\ Z_s &= \mu_s + m_s \end{aligned}$$

where ϵ_s is measurement error. In a purely spatial analysis, repeated observations at a specific location over time are treated as independent realisations of the underlying process. The observed data y_s at the first level of the model are considered conditionally independent given the value of the underlying process.

The concept of stationarity is critical to most of spatial analyses. A stationary process has no spatial trend, meaning a constant mean, with the covariance between two points dependent only on the distance between them and not their actual locations. If a process is stationary, measurements from any area within the study region can be used to make inference about the overall underlying structure. A spatial process is strictly stationary if the joint distribution is invariant in space, meaning $f(Z_{s_1}, \ldots, Z_{s_{N_s}})$ is the same as $f(Z_{s_1+h}, \ldots, Z_{s_{N_s}+h})$ for any N_s and distance $h \in \mathcal{R}^d$. However this criteria is excessively restrictive and is rarely achieved in practice. Two weaker and most common stationary assumptions are second order stationary (weak stationary) and intrinsic stationary. A spatial process is second order stationary if for any locations s and s^*:

$$\begin{aligned} E[Z_s] &= E[Z_{s^*}] \\ Cov[Z_s, Z_{s^*}] &= \psi(h) \end{aligned}$$

where $h \in \mathcal{R}^d$ is the distance between two locations s and s^*. This implies that the variance is constant over the entire region and that the covariance between two locations depends only on the distance, h, and not direction between them, known as isotropy.

A further simplification is *intrinsic* stationary, which is based on the difference between two locations. The difference in means is zero and the difference in variance is defined through the semi-variogram.

$$\frac{1}{2} Var[z_s - z_{s^*}] = \psi(\|h\|)$$

where $\|.\|$ denotes the Euclidean norm, meaning the variance is constant over the entire region. This implies that the variance of the difference must be the same everywhere in the region, but does not require that the variance itself of observations is constant over the entire region.

The covariance function and the semi-variogram are both functions that summarize the strength of association between responses as a function of distance, and possibly direction. In practice, the semi-variogram is often preferred to the covariance function because of the relaxed rules of stationarity and also because it only uses pairs of locations h units apart, and does not involve the overall mean, so if there is a shift in the mean which is not explicitly modelled, the semi-variogram is likely to be less affected than the covariance function. A common class of spatial models is the Matérn class

$$Cov[Z_s, Z_{s^*}] \;=\; \frac{\sigma^2}{\Gamma(\nu)2^{\nu-1}}\big(\kappa \parallel h \parallel\big)^{\nu} K_{\nu}\big(\kappa \parallel h \parallel\big)$$

where σ^2 is a scalable parameter controlling the range of the spatial correlation, it is defined by

$$\sigma^2 \;=\; \frac{\Gamma(\nu)}{\Gamma(\nu + d/2)(4\pi)^{d/2}\kappa^{2\nu}}$$

here d is the number of dimensions from $\mathcal{R}^d$, K_ν is the modified Bessel function of the second kind with order ν which is usually fixed and determines differentiability of the sample paths. The relation between scaling parameter κ and range parameter ρ is empirically derived as $\rho \;=\; \sqrt{8\nu}/\kappa$. There are two special cases of Matérn class models

- Gaussian $\qquad \psi(\parallel h \parallel) \;=\; \sigma^2 exp(-\theta \parallel h \parallel^2)$
- Exponential $\qquad \psi(\parallel h \parallel) \;=\; \sigma^2 exp(-\theta \parallel h \parallel)$

The exponential model is a special case with $\nu = \frac{1}{2}$, while the limiting case of the Matern class, when $\nu \to \infty$, is the Gaussian model.

2.2 Temporal processes

Temporal stochastic processes are used to model both overall temporal trend and temporal correlations with data. Classical time series composition and analysis can be very useful in understanding the nature of any serial dependence and thus in constructing suitable models (Chatfield, 2013).

The main aims of classical time series analysis are description, modelling, forecasting and control. In this thesis, the main interest is modelling where the response can depend on past or present values of other explanatory variables. The classical time series modelling aims to decompose the variation in the series into four components. The first one is trend that describes the long term movements in the mean. The second element is seasonality which represents annual cyclical fluctuations. There are also other cyclical variations, at frequencies less than or greater than a year. The last component is the residual which represents other random or systematic fluctuations. After modelling these four elements, there may still be autocorrelation in the residual term, and it is this fourth component that is the one of interest in the context of this thesis.

A time series is said to be stationary if the distributional structure of outcome Y_t is unaffected by a shift in time. Strict stationary means that for any choice of k, and times $t, t = 1, \ldots, N_t$, the joint probability distribution of $Y_{t+k}, t = 1, \ldots, n$ is the same for all k. i.e. the joint distribution does not depend on the location in time. So, if a process is stationary, observations from any time period can be used to make inference about the overall underlying structure. A weaker assumption is one of weak or second order stationarity where the mean, $E(Y_t) = \mu_t$ is constant for all t, and the auto-covariance function $Cov(Y_k, Y_t)$ or $Cov(Y_k - \mu_k, Y_t - \mu_t)$ depends only on the distance (in time) between k and t and not their actual location. Time series models are often of the form

$$Y_t = \mu_t + \epsilon_t$$

where ϵ_t is a stationary random function, which is often referred to as 'noise', and μ_t is the trend. In order to use a model of this type, assumptions have to be made about the form of the trend and the noise functions. If the trend can be modelled successfully, then the noise term may be considered to be stationary.

2.3 Spatio-temporal modelling

A spatial-temporal random field, $Z_{st}, s \in \mathcal{S}, t \in \mathcal{T}$, is a stochastic process over a region and time period. This underlying process is not directly measurable, but realisations of it can be obtained by taking measurements, possibly with error, at a set of known locations in space $S = \{s_1, ..., s_{N_S}\} \in \mathcal{S}$ and time $T = \{t_1, ..., t_{N_T}\} \in \mathcal{T}$. In a purely spatial analysis, repeated observations at a specific location over time are treated as independent realisations of the underlying process. There are three levels to the hierarchy that we consider. The observed data, $Y_{st}, s = 1, ..., N_S, t = 1, ..., N_T$, at the first level of the model are considered conditionally independent given a realization of the underlying process, Z_{st}. The second level describes the true underlying process as a combination of a trend (mean), μ_{st}, and a random process, ω_{st}, which has spatial–temporal structure in its covariance. In a Bayesian analysis, the third level of the model assigns prior distributions to the hyperparameters from the previous levels. Thus in summary we have:

$$\begin{aligned} Y_{ts} &= Z_{ts} + \epsilon_{ts} \\ Z_{ts} &= \mu_{ts} + \omega_{st} \end{aligned} \tag{2.1}$$

where the $\{\epsilon_{ts}\}$ is an independent random, or measurement, error term, μ_{ts} is a space-time

mean field (trend) and ω_{st} is a spatial–temporal process.

Models in which the spatial and temporal structure are modelled jointly are known as *non-separable* models. Alternatively, a simpler approach is *separable* models, where the temporal and spatial components are modelled separately without any interaction. These models impose independence between space and time components, which, although often an overly simplistic assumption, can lead to a reduction in computation. Given two measurements, Y_{st} and $Y_{s^*t^*}$, made in time and space let ψ_{ss^*} be the correlation between them in space and ψ_{tt^*} the correlation over time. Separability can either be multiplicative or additive. In the additive case, the covariance is separable, it can be written as the sum of ψ_{ss^*} and ψ_{tt^*}:

$$Cov[Y_{st}, Y_{s^*t^*}] = \psi_{tt^*} + \psi_{ss^*}$$

Separable covariance functions are widely used in the air pollution literature, for example Shaddick and Wakefield (2002) considered the spatio-temporal modelling of four pollutants measured daily at eight monitoring sites in London. They modelled the data within a dynamic linear modelling framework, using a Bayesian approach with implementation via MCMC. Sahu et al. (2006) used a separable spatio-temporal model for fine particulate matter in three midwestern U.S. states.

2.4 Spatio-temporal models in environmental applications

Zidek et al. (2002) implemented an approach suggested by Le et al. (1997) to model the space-time field of daily ambient PM10 in Vancouver, Canada. For simplicity, they analysed each monitoring site separately and chose an AR(1) model to represent the temporal structure. They identified the possibility that spatial correlation between sites might 'leak' into the lagged values of the series, due to modelling each site univariately. This would not have happened if a non separable spatio–temporal model could have been used, but such an approach may be infeasible with the large number of monitored and unmonitored sites.

In an early Bayesian application, Handcock and Wallis (1994) used a spatio-temporal approach for winter temperatures. Their approach was to carry out separate spatial analyses in each year. The mean and covariance parameters of these models were then examined and found to be stable over time. They then assumed that spatial and temporal aspects could be modelled separately.

More recently, Sahu et al. (2007) developed a spatio-temporal model for the analysis of daily ozone observations in Ohio, U.S. They used the square root of the observations and fitted

a stationary auto-regressive model for the temporal correlation and an exponential covariance structure for the spatial correlation. Vanem et al. (2012) used a Bayesian hierarchical spatio-temporal model for wave height in the North Atlantic. Cameletti et al. (2011) used a separable spatio-temporal model to analyse PM10 data in the Po valley (northern Italy). The authors compared 6 types of spatio-temporal model which feature different levels of complexity either in the hierarchical structure or in the spatio-temporal covariance function. The comparison is based on criteria that take into account intrinsic complexity, computational costs and spatial prediction capability. Jürgens et al. (2013) explore age and gender specific spatio-temporal patterns of lung cancer and other tobacco-related cancer mortality rates in Switzerland using Bayesian hierarchical spatio-temporal models. They use this approach to determine differences between rural or urban living, and extend the study into age and gender specific mortality by linguistic region.

2.5 Measurement error

Measurement error is a general term used to encompass situations where the observed data do not represent the quantity of interest exactly. It can occur in both response variables and covariates. Epidemiological studies require accurate measurements of both health outcomes exposures together with potential confounders. All of these data may be measured with varying degrees of error. Considering measurements of air pollution; there is a true underlying pollution surface which will drive the exposures experienced by the population at risk. However this surface is not directly observable and instead measurements are taken at locations over space and time. Differences between these exposure measurements and the unknown underlying field is often referred to as measurement error. Here the term measurement error is taken to refer to any difference from the underlying true values (of pollution) and what is measured.

Traditionally measurement error has been based around the idea of repeated measurements of a value, for example measuring blood pressure, in which the repeated measurements will contain a component of error; often assumed to be random. In spatial epidemiology, error will possibly comprise of a number of factors including monitor calibration error and random variation but also variations in the underlying pollution field over time and space which are not acknowledged in the model. This may arise for example when the modelling assumptions are too simplistic for the complex surface of the pollution field. Another issue that is often contained under the umbrella term of 'measurement error' is the misalignment of locations or times of exposure measurements and health outcomes. This arises because exposure and health data are often drawn from independent sources and not as the result of a carefully designed study. Hence a straightforward comparison is not possible without a model to align these elements in

the spatial and temporal domains (Gryparis et al., 2009; Peng and Bell, 2010). In such settings, health effects analysis may use predictions from an exposure model as substitutes for actual measures of exposures in the health model.

A brief review of measurement error models is given here, more comprehensive discussions are given by Fuller (1987) and Carroll et al. (2006), which focus on linear and non-linear models respectively. Measurement error models are based on four quantities;

- Z - the true unobserved exposures ;
- Y - the observed exposures which are measurements of Z, potentially incorporating some measure of error;
- X - covariates, which are assumed to be measured exactly.

The joint likelihood of these quantities can expressed as $f(Z, Y|X)$, where the covariates are conditioned on because they are fixed and known. This represents the relationship between the unobserved exposure Z and the measured surrogate Y. There are two types of measurement error model; classical and Berkson which are outlined below.

2.5.1 Classical measurement error

Classical measurement error models decompose $f(Z, Y|X)$ into $f(Y|Z, X)f(Z|X)$, the first element is a conditional model for the measured surrogate Y given the true (unobserved) exposure Z. Two common classical measurement error models are (i) additive and (ii) error calibration.

(i) $Y_i \sim N(Z_i, \sigma^2)$ for $i = 1, ..., n$

(ii) $Y_i \sim N(\beta_0 + \beta_z Z_i + \sum_{p=1}^{P} \beta_p X_{pi}, \sigma^2)$ for $i = 1, ..., n$

In the simple additive formulation the observed surrogate is assumed to be correct on average (that is $E[Y_i|Z_i] = Z_i$), while in model (ii) the surrogate is biased. Both models specify an additive relationship between Y_i and Z_i, an alternative being a multiplicative error model $Y_i = Z_i\epsilon_i$, where ϵ_i is a zero mean Gaussian error with variance σ^2. The remaining term $f(Z|X)$ can be based on knowledge of thet true exposure or represent prior ignorance. In a Bayesian setting $f(Z|X)$ acts as a prior for the unknown exposure Z.

2.5.2 Berkson measurement error

In contrast to the classical case, Berkson measurement error models decompose $f(Z, Y|X)$ into $f(Z|Y, X)f(Y|X)$, where the first term is a conditional model for the true exposure Z given the measured surrogate Y. Again there are two common approaches; (iii) additive and (iv) regression calibration. In common with the classical models $f(Z|Y, X)$ is a decomposition of of independent distributions for each observation.

(iii) $Z_i \sim N(Y, \sigma^2)$ for $i = 1, ..., n$

(iv) $Z_i \sim N(\beta_0 Y_i + \beta_Y Y_i + \sum_{p=1}^{P} \beta_p X_{pi}, \sigma^2)$ for $i = 1, ..., n$

In the simple additive model the true exposure is assumed to be equal to the surrogate on average (that is $E[Z_i|Y_i] = Y_i$), but this is not true for (iv). As with classical models a multiplicative alternative can be used, which is implemented using an additive model on the log scale. In the Berkson model, as Y are known measurements, the distribution $f(Y|X)$ can be ignored. The choice between classical and Berkson models will depend on the structure of the problem as well as the set of available data. Further details can be found in Carroll et al. (2006).

The measurement error models described above can only be used if additional data are available, because the information from (Y, X) is not sufficient to estimate the measurement error process. Examples of such additional data include repeated measurements of Y which in a spatial setting may be measurements at each location over time. Alternatively, external data may be able to inform the process if observed values of Z and Y were available at a subset of locations. The identifiability of a proposed model may also depend on the assumptions made about the measurement error process. There are two generic classes of such assumptions; functional and structural. Functional models are distribution invariant and specify minimal assumptions about the measurement error process. They do not specify a proper likelihood, and estimation is typically based on regression calibration. In contrast structural models, such as those shown in (i) to (iv), are fully parametric and specify probability distributions for $f(Y|Z, X)$ or $f(Z|Y, X)$. The choice between functional and structural models determines the method of estimation and inference that can be used, with structural models enabling likelihood and Bayesian methods to be applied. More details can be found in Carroll et al. (2006).

Chapter 3

Modelling health risks

Potential associations between exposure to air pollution and mortality (or morbidity) are the main focus of this thesis. The strength of an association is usually expressed in terms of the relative, or change in, risk (RR) associated with a change in air pollution. In the environmental settings considered in this thesis, the relative risk is commonly estimated from ecological data using Poisson log–linear models. These data typically comprise area level summaries of mortality or morbidity, ambient pollution levels at fixed locations and meteorological covariates. The nature of the data presents a number of statistical challenges, including unmeasured confounding, as the associations of interest are typically small. This means that estimation can be difficult and accurate and realistic models are important. However, as models become more realistic they may increase in complexity, requiring more data and computational power to estimate parameters. Often therefore the choice of statistical model results in a trade–off between simple models that are computationally efficient and easy to interpret, and more complex alternatives which make less unrealistic assumptions about the data but may be more difficult to fit and may often offer less suitable interpretation.

In this chapter, the standard approaches to modelling health data and related factors are outlined. The first section gives an introduction to epidemiology and the idea of relative risk. Section 3.2 describes the type of health data that is often used in epidemiological studies of air pollution and health and how they are collected. Section 3.3 outlines the covariate risk factors that may influence the relationship between health data and exposure data, including both known and unknown risk factors. Regression models and principles of inference are presented in Section 3.4, in which generalised additive models (GAM), which are widely used for health modelling in this setting, are introduced. In Section 3.5, splines are introduced as an approach to modelling temporal patterns in health analyses, and penalized splines are outlined. The last section describes the issue of over dispersion which is commonly seen in health studies, and

the quasi-likelihood approach which may be used to acknowledge it in estimating measures of uncertainty.

3.1 Epidemiology and relative risk

Epidemiology is defined as the study of factors that determine the occurrence and distribution of disease in a population (Jekel et al., 2007). In epidemiology, risk is defined as the proportion of the population at risk who are unaffected at the beginning of a study period, but who undergo the risk event during the study period.

The real interest in our research is the estimation of relative risk (RR), which is also known as the risk ratio. The RR is the ratio of the risk in an exposed group compared to that in an unexposed group. If the risks in the exposed group and unexposed group are the same, then RR = 1. If the risks in the two groups are not the same, calculating the RR provides a way of showing in relative terms how much different (greater or smaller) the risk in the exposed group is compared with the risk in the unexposed groups. The risk for the disease in the exposed group usually is greater if an exposure is harmful, for example cigarette smoking; or smaller if an exposure is protective, as in the case of a vaccine. It also is important to consider the number of people to whom the relative risk applies. A large relative risk that applies to a small number of people may produce few excess deaths or cases of disease, whereas a small relative risk that applies to a large number of people may produce many excess deaths or cases of disease. In the case of air pollution, the RR may be defined between for example cities that have high levels of air pollution (exposed) and those with low levels (unexposed) or between days with high levels (exposed) and those with low (unexposed). Often RRs in these cases are expressed in terms of changes in levels of pollution, for example per $10\mu gm^{-3}$ increase in particulate matter in which case the relative risk is

$$RR = \frac{predicted\ health\ count\ at\ exposure\ (A+10)}{predicted\ health\ count\ at\ expoure\ A} \tag{3.1}$$

In the following sections, we consider the type of health data and information on potential confounders that commonly occur in studies estimating the RR associated with air pollution.

3.2 Health data

Mortality or morbidity data are often only available as aggregated daily counts within a geographical region of study. They may comprise the number of mortality or morbidity events occurring each day. They are collected from hospital records and death registries, and for confidentiality reasons are not available at the individual level. All mortality events are classified

by cause of death using the international classification of disease (ICD) (WHO, 1980). A variety of mortality classifications have been used in studies of air pollution and health, the most general of which is total non-accidental mortality. However this includes a significant proportion of deaths that are unrelated to pollution exposure, which may cause the pollution-mortality association to be biased. Consequently, cause specific outcomes such as mortality due to respiratory or cardiovascular illness may be preferable, because they are more likely to be related to the possible effects of air pollution. However, using a more precise definition for the response may result in smaller numbers of mortality events and no more accurate estimation of associations with pollution. Many studies have also analysed mortality data relating to specific age groups such as the elderly or children, because these frail sub-populations are more likely to be susceptible to air pollution than the general population. In addition to mortality the association between air pollution exposure and morbidity has also be investigated, with positive associations found for asthma and respiratory and circulatory illness (Schwartz, 2001; Pope et al., 1995).

3.3 Confounders

In addition to the possible effects of air pollution, counts of mortality or morbidity will depend on a set of other risk factors, and if the influence of these factors is not adequately removed, then the estimated pollution-mortality association may be biased. This is known as confounding. This may induce long-term trends, seasonal variation, over-dispersion and temporal correlation into the health data. Confounders may include meteorological conditions such as temperature, humidity, wind speed, and rainfall. In addition to measured confounders there may also be effects of unmeasured factors. These might be represented by proxy variables such as functions of calender time and variables that indicate the day of the week.

3.3.1 Known risk factors

Measured risk factors are typically related to meteorological events. Meteorological covariates often include temperature (Mar et al., 2000), humidity (Lee et al., 2000), precipitation (Spix et al., 1993), and pressure (Vedal et al., 2003). The most common of these is temperature, because it causes part of the seasonal variation typically present in health data, for example, higher counts of mortality during cold period. Although meteorological covariates are routinely available their inclusion in an epidemiological model requires a number of decisions including which lag should be used and what shape should its relationship with health take.

The health problems that result from pollution exposure may be felt immediately, that is on the same day (Moolgavkar, 2000), after a lag of one or two days (Peters et al., 2000), or

from continued exposure over the preceding few days or from long–term exposure over several decades (Elliott et al., 2007). The choice between different lags is a longstanding research problem, and there is no consensus over which should be used. It has been determined by numerous approaches, including selecting the lag that maximizes the estimated adverse effect (Lumley and Sheppard, 2000), the one used by previous studies, or the one that minimises an objective criteria (such as DIC). Alternatively, results for multiple lags have been presented, for example (Burnett et al., 1994).

3.3.2 Unknown risk factors

Unknown risk factors may result in long-term trends and seasonal variation in time series studies and large-scale spatial trends, for example north to south gradients, in spatial studies. They can not be added to regression models in the same way as known factors. Removing the influence of unknown risk factors is less straightforward. In early temporal studies, Schwartz et al. (1993) and Spix et al. (1993) modelled seasonal variation with pairs of sine and cosine terms at different frequencies, and long-term trends with parametric function cubic polynomials of calender time, such as quadratic. Other early approaches model these factors with indicator variables (Verhoeff et al., 1996) which like the parametric functions described may be overly restrictive and lack the necessary flexibility to model excessive variation in mortality. For example the sinusoidal terms force the peak in mortality to occur at the same time each year, while the monthly indicator variables do not allow for within month variation. Nowadays, these unmeasured risk factors are represented using smooth functions of calender time, which can be more flexible than fixed parametric alternatives. Such functions have been implemented using parametric and non-parametric methods, including regression splines space (Daniels et al., 2004), smoothing splines (Dominici et al., 2000).

Categorical, or indicator, variables are often used as proxies for factors that may confound the relationship of interest. These include ‘day of the week’ (see for example Kelsall et al. (1999)), influenza epidemics (Peters et al., 2000) and public holidays (Schwartz, 2001). In spatial studies, where the RR is driven by differences in health counts between different areas, confounding variables might for example represent the effects of socio-economic deprivation which has been shown to be a strong predictor of both health (Kleinschmidt et al., 1995) and air pollution (Elliott et al., 2007).

3.4 Regression modelling

Fitting a regression model generally involves explicitly stating the form of the relationship between the explanatory and response variables being examined, hence makeing clear the as-

sumptions that underpin the analysis. However, this may be difficult to understand when complex statistical methods are used and it is important to express the results in a form that can be easily interpreted and understood by both statisticians and epidemiologists. The model fitting requires many aspects for consideration, including the choice of distribution, for instance Normal, Poisson or Binomial distributions, the choice how associations will be modelled, for example, linear, log-linear or logistic regression, and how the actual computation is carried out.

The main focus of the use of regression models in this thesis, and most common reason for performing a regression analysis in epidemiology, is to obtain estimates of the coefficients associated with the variable of interest, for example, the effect of an increase in air pollution on the risk of an adverse health outcome.

We now describe two general frameworks for regression modelling used within this thesis: generalised linear models (GLM) and generalised additive models (GAM).

3.4.1 Generalised linear models

Generalised linear models (GLMs) are extensions of linear models (McCullagh and Nelder, 1989). In a GLM, the response variable is assumed to be an independent observation from an exponential family distribution and is related to the exposure variables and covariates through a link function. If the response variables are denoted by $Y = (Y_1, \ldots, Y_n)$, and q covariates represented by $X = (X_1, \ldots, X_n)_{n \times q}$, then the general form of generalised linear model is given as

$$\begin{aligned} Y_i &\sim f(Y_i|\mu_i) \quad for \;\; i = 1, \ldots, n, \\ g(\mu_i) &= X_i\theta \end{aligned} \tag{3.2}$$

where μ_i denotes the expected value of Y_i. The $\theta = (\theta_1, \ldots, \theta_q)$ are the unknown regression parameters which represent the relationship between the explanatory variable and the response. The linear combination of all the covariates is called the linear predictor, and is related to the expected value μ_i, via an invertible link function g. The unknown θ can be estimated using both likelihood and Bayesian methods. For the likelihood method, the parameters θ can be estimated by solving the estimating equations:

$$\sum_{i=1}^{p} \frac{d\mu_i}{d\beta} v_i^{-1}(Y_i - \mu_i(\theta)) = 0$$

where $v_i = VAR(Y)$. In the Bayesian case, prior distributions need to be assigned to all the

parameters.

In the vast majority of epidemiological studies using aggregate level data, either over time or space or both, the response variables are assumed to follow a Poisson distribution. Counts (of health outcomes) will be within the range $[0, \infty)$. The natural link function is therefore log, which restricts the outcome to the required range. It is noted that this implies an underlying multiplicative relationship between the effects of the covariates.

3.4.2 Generalised additive models

Genaralised additive models (GAM) can be considered as extensions of genralised linear models, where instead of assuming dependence on the sum of linear predictors, the outcome is assumed to be dependent on a sum of functions of the predictors. GAMs therefore provide a flexible framework for controlling for non-linear dependence on both the variable of interest and potential covariates. Assuming Y_i to be an independent observation from an exponential family distribution f, then the model is given by

$$\begin{aligned} Y_i &\sim f(Y_i|\mu_i) \quad for \quad i = 1, \ldots, n, \\ g(\mu_i) &= \beta_0 + \sum_{j=1}^{q} S_j(X_{ij}|\lambda_j) \end{aligned}$$

where the relationship between $g(\mu_i)$ and each covariate x_{ij} is represented by a function, S_j. The functions S_j can be estimated using a number of methods, two widely used approaches are kernel smoothing (Hastie and Tibshirani, 1990) and smoothing splines (Brezger and Lang, 2006). The smoothness of each function is controlled by the parameter λ_j.

3.5 Splines

Splines are a flexible technique for modelling non–linear relationships. They transform a possible non–linear relationship into a linear form by separating the data x into $k + 1$ sub-intervals. The points where sub-intervals join are known as the knots of the spline, denoted by $k_1, \ldots, k_K$. Consider a smooth function $f(x)$ of the form:

$$f(x) = \sum_{i=1}^{q} \beta_i b_i(x)$$

where $b_i(x)$ is the i^{th} basis function and β_i is the basis parameter. There are many choices of basis function to represent the spline, for example, polynomial spline, penalized regression

spline, B-spline and radial basis spline (Wood, 2006).

3.5.1 Penalised splines

A penalised spline uses an overly large number of basis functions and penalises excess curvature by using a penalty term. The smoothness of the splines depends on the number of knots: too many knots may lead to over smoothness, whilst inadequate number of knots leads to rough model fit. The key to penalised splines is controlling smoothness by adding a 'wiggliness' penalty to the least squares objective, that is fitting the model by minimizing:

$$\|Y - X\beta\|^2 \; + \; \lambda \int_a^b \left[f''(x)\right]^2 dx$$

where the second part of this formula penalizes models that are too 'wiggly'. The trade off between model fit and model smoothness is controlled by the smoothing parameter, λ. Since f is linear in the parameters β_i, the penalty term can be written as a quadratic form in β:

$$\int_a^b \left[f''(x)\right]^2 dx = \beta^T S \beta$$

where S is a matrix of known coefficients. Therefore, the penalized regression spline fitting problem is equivalent to minimizing:

$$\|Y - X\beta\|^2 + \lambda \beta^T S \beta$$

In this case, the estimation of degree of smoothness becomes the issue of selecting smoothing parameter λ. If λ is too high then the data will be over smoothed, and if it is too low then the data will be under smoothed, in both cases this will mean that the spline estimate $\hat{f}$ will not be close to the true function f. Choosing λ may be done using data driven criterion, such as cross validation (CV) and generalised cross validation (GCV), details of which can be found in Wahba (1990) and Gu (2013).

3.6 Over-dispersion

If all risk factors that influence health are known and included in the regression model, the residual variation would be adequately described by the Poisson assumption. However it is likely that only a subset of these risk factors are known, and the presence of unknown factors may inflate the variance and make the Poisson assumption untenable. This will cause confidence intervals to be too narrow, which in some cases may falsely suggest that the pollution-health association is statistically significant. Increased variation in the data compared with that specified by the assumed probability distribution is known as over-dispersion. The unmeasured

confounders may be operating at the individual level, e.g. smoking, or at the area level, e.g. residual socio-economic confounding. Over-dispersion may also arise because of data anomalies such as errors in the numerators and/or denominators or due to migration.

The choice of a log link and variance that is proportional to the mean is the canonical one for the Poisson distribution and under regular Poisson assumptions, the variance is assumed to be equal to the mean, $Var(Y) = E(Y)$. When this is relaxed to proportionality, $Var(Y) = \phi E(Y)$, the excess variation is modelled by the dispersion parameter ϕ which is assumed to be constant over all of the data. The estimating equations for this will generally be different from those obtained by weighted least squares, but solutions can be found using quasi-likelihood techniques.

Quasi-likelihood only requires that the first two moments of the data generating distribution are specified, i.e. the mean and variance. The quasi-likelihood is derived by approximating the score function for Y with $\frac{Y_i - k}{\phi V(k)}$. Incorporating over-dispersion scales the variance of the estimates by the estimated dispersion parameter

$$\widehat{\phi} = \frac{1}{n-q} \sum_{i=1}^{n} \frac{(Y_i - \mu_i)^2}{V(\mu_i)}$$

which widens the corresponding confidence intervals. Although a standard technique in likelihood based inference, quasi-likelihood methods can not be applied in a Bayesian setting because a complete likelihood is required. A comprehensive description of quasi-likelihood is given by McCullagh and Nelder (1989).

Chapter 4

Bayesian analysis

The Bayesian approach provides a natural framework for dealing with hierarchical models, incorporating the uncertainty that will be present at each stage of the process in a coherent manner and acknowledging it in the estimates of the parameters of interest. In this chapter, we give a brief review of Bayesian analysis. It starts with an introduction to Bayesian inference, including a description of Bayes theorem, prior distributions and posterior distributions. We then describe the differences between a fully Bayesian, or *one–stage* approach and *two–stage* approaches in which different parts of the overall model may be fitted separately.

4.1 Bayesian inference

The aim of Bayesian inference is to make inference about parameters, θ, given observed data Y. In order to achieve this, a joint probability distribution relating the parameters θ and the observed values Y is expressed as a product of two densities; the prior distribution $f(\theta)$ and the likelihood function of Y given the values of parameters, $f(Y|\theta)$;

$$f(\theta, Y) = f(Y|\theta)f(\theta) \tag{4.1}$$

Applying Bayes' theorem allows to express the posterior distribution of the parameters given the data, $f(\theta|Y)$, this represents the updated beliefs about parameters θ after the information of data is taken into account,

$$f(\theta|Y) = \frac{f(\theta, Y)}{f(Y)} = \frac{f(Y|\theta)f(\theta)}{f(Y)} \tag{4.2}$$

where the denominator $f(Y)$ is the marginal distribution of the data, and is calculated as $f(Y) = \sum_\theta f(\theta)f(Y|\theta)$ if θ is discrete, and $f(Y) = \int_\theta f(\theta)f(Y|\theta)d\theta$, where the integration is over all possible values of θ, if θ is continuous. The denominator $f(Y)$ which does not

depend on any of the unknown quantities, θ, can be treated as a normalising constant. Hence equation (4.2) can be simplified to:

$$f(\theta|Y) \propto f(Y|\theta)f(\theta) \tag{4.3}$$

which shows that the posterior is the product of the likelihood function and the prior distribution. The ability to summarise parameters via their posterior distribution is a one of the significant features of Bayesian analyses, because $f(\theta|Y)$ contains more information about θ than is typically obtained from a frequentist analysis based on likelihood inference which commonly results in just point estimates.

The ability to summarise θ via its posterior distribution is a major advantage of the Bayesian framework, because $f(\theta|Y)$ contains more information about θ than is typically obtained from a frequentist analysis. In Bayesian statistics, the posterior mean and median are typically quoted as point estimates, while a credible interval is an interval in the domain of a posterior probability distribution used for interval estimation, meaning that a $100(1-\alpha)\%$ credible interval is an interval within which $100(1-\alpha)\%$ of the posterior distribution lies. We need to notice that the credible interval is different from the concept of confidence interval used in frequentist statistics, where confidence interval gives an estimated range of values which is likely to include an unknown population parameter, meaning that for a $100(1-\alpha)\%$ confidence interval, we are $100(1-\alpha)\%$ confident that the true value of the parameter is in our confidence interval. The reminder of this section discussed prior and posterior distributions, more general review of Bayesian methods is given by Gelman (2003).

4.1.1 Prior distributions

The prior distribution represents the belief about θ before any data are observed. It allows knowledge from previous studies or experiments to be incorporated. The choice of prior distribution will depend on the context, but is typically represented by a standard probability distribution which depends on a vector of hyper-parameters that may or may not be known. If there are doubts as to the accuracy of the prior distribution, it is important to assess the sensitivity of the posterior probabilities to the choice of priors.

A conjugate prior is one that has the distribution in the same family as the posterior distribution. For example consider a Gaussian density function for the observation $y \sim N(\mu, \sigma^2)$ with known parameter μ and unknown parameters σ^2. The conjugate prior of σ^2 is $Inverse-Gamma(a,b)$ which results in the following full conditional distribution:

$$f(\sigma^2|\mu, y) \sim Inverse - Gamma(a + \frac{1}{2}\ ,\ b + \frac{1}{2}(y-\mu)^2) \tag{4.4}$$

The advantage of using conjugate priors is that it makes the computation of the posterior distribution relatively straightforward. Despite the convenience that using a conjugate prior provides, there will be occasions when they will not be a suitable specification of prior knowledge, and therefore would not be an appropriate choice. Alternatively prior ignorance, meaning there is only vague information about a quantity, can be represented by using non-informative prior which gives fairly even support to a wide range of values. Non-informative priors can be a standard distribution with a large variance or may be improper. An improper prior is one such that $\int f(\theta)d\theta = \infty$, it should be used with caution, as the resulting posterior distribution may also be improper. As a result, prior ignorance is often specified using a proper prior with large variance, with two common choices being Gaussian, generally used for mean or regression parameters, or inverse-gamma distributions, for variances.

4.1.2 Posterior distribution

The posterior distribution can be calculated using many approaches, and the choice of which will be largely determined by the form of $f(\theta|Y)$. For simple situations with conjugate priors, since the posterior distribution comes from a standard family of distributions, it can be obtained analytically (see, for example, Gelman (2003)). However in most cases this is not possible, as such conjugacy is either not available or overly restrictive.

One approach to finding posterior probabilities is to use simulation techniques, such as Monte Carlo. Monte Carlo integration evaluates the population mean by a sample mean. In this situation a large number of samples are drawn from $f(\theta|Y)$ to estimate quantities of interest such as the posterior mean and median, as well as credible intervals. When the samples are independent, laws of large numbers ensure that the approximation can be made as accurate as desired by increasing the sample size. The samples can be generated using a number of methods, the simplest of which are direct approaches such as inversion or rejection sampling (Gamerman and Lopes, 2006). In general, drawing samples independently from the posterior distribution is not feasible, since it can be quite non-standard. However the samples can be generated by any process throughout the support of the posterior distribution in the correct proportions. One way of doing this is Markov chain monte carlo (MCMC) simulation, for an introduction to MCMC, see Gilks et al. (1996), and Gamerman and Lopes (2006). There is a difference between MCMC and Monte Carlo simulation, as simple Monte Carlo simulation simulates independent random values from the probability distribution of interest, but there is dependence between simulated samples in MCMC methods. Details of the MCMC used in this

thesis can be found in Chapter 7.

4.2 Linking health and exposure models

In a fully Bayesian analysis, estimation for both the health and exposure models, including prediction at locations where data are not available, would be performed simultaneously. The uncertainty in estimating the coefficients of the exposure model is therefore acknowledged and 'fed through' the model to the predictions and thus to the estimation of the coefficients in the health model.

There are likely to be computational considerations associated with jointly fitting the health and exposure models, especially if the latter uses large amounts of data over space and time. When the exposure model is complicated or when one is interested in running multiple candidate epidemiological models with different sets of covariates either for a single outcome or multiple outcomes, a single model is not going to provide an efficient method of investigation.

Often the exposure models are fitted separately from the health model, removing the dependence of $Z^{(2)}$ on $Y^{(1)}$. The joint model is therefore decomposed into separate health and exposure components. The exposure component is of the form:

$$f(Z^{(2)}|\theta^{(2)}, Y^{(2)}) \propto f(Y^{(2)}|Z^{(2)}, \theta^{(2)}) f(\theta^{(2)})$$

and the health component of the form

$$f(\theta^{(1)}|Z^{(2)}, Y^{(1)}) \propto f(Y^{(1)}|Z^{(2)}, \theta^{(1)}) f(\theta^{(1)})$$

noting that the first term exposure model is different from the $f(Z^{(2)}|\theta^{(2)}, Y^{(2)}, Y^{(1)})$, that would be the case in a fully Bayesian analysis. This is often done in order to ease the computational burden in running a combined model, and that has been adopted in a number of cases, for example, Carlin et al. (1999), Zhu et al. (2003), Lee and Shaddick (2010), Chang et al. (2011) and Peng and Bell (2010). This *two-stage approach* has the advantage that the exposure model, which is likely to be the most computationally demanding, does not have to be refitted when running multiple health effect analyses. Two stage approaches separate the exposure and health components, whilst still allowing uncertainty from the exposure modelling to be incorporated into the health model.

There are other reasons why fitting a joint model may be unappealing; it is not intended that the health counts should inform the estimation of the exposures which should be based

on data from the monitored concentrations. It is possible to 'cut' feedback between the stages within MCMC, for example in WinBUGS (Lunn et al., 2000), however the result is that the posteriors may not be proper probability distributions.

4.2.1 Two-stage Bayesian approaches

Peng and Bell (2010) developed a two-stage approach involving two separate Markov chain Monte Carlo implementations, they firstly estimate $f(Z_t^{(2)}|Y_t^{(2)})$, which is the posterior distribution of exposure $Z_t^{(2)}$ given the observation $Y_t^{(2)}$ for each time t. The second stage uses $f(Z_t^{(2)}|Y_t^{(2)})$ as an informative prior for $Z_t^{(2)}$ and estimates the joint posterior distribution of θ and $Z_t^{(2)}$, given the health data $Y_t^{(1)}$ and the observed pollution data $Y_t^{(2)}$,

$$f(Z^{(2)}, \theta|Y^{(1)}, Y^{(2)}) \propto f(Y^{(1)}|\theta, Z^{(2)}, Y^{(2)}) f(Z^{(2)}|Y^{(2)}) f(\theta)$$

where $f(\theta)$ is a diffuse prior distribution. The likelihood terms $f(Y^{(1)}|\theta, Z^{(2)}, Y^{(2)})$ represent the Poisson likelihood used for health model. In fact, only the second stage of the model is Bayesian, while the first stage is estimated with maximum likelihood and is treated as a fixed prior in the second stage. This two-stage approach effectively assumes that $f(Z^{(2)}|Y^{(1)}, Y^{(2)}) \approx f(Z^{(2)}|Y^{(2)})$, thus cutting the feedback between $Z^{(2)}$ and $Y^{(1)}$. Chang et al. (2011) use another two-stage framework, for the first stage, posterior samples of exposures given the measured data are obtained by sampling from the following posterior predictive distribution:

$$Stage1: \quad f(Z^{(2)}|Y^{(2)}) \propto \int f(Y^{(2)}|Z^{(2)}, \theta_1) f(Z^{(2)}|\theta_2) f(\theta_1, \theta_2)\, d\theta_1 d\theta_2$$

where $f(Y^{(2)}|Z^{(2)}, \theta_1)$ represents the measurement model and $f(Z^{(2)}|\theta_2)$ represents the exposure model. Here the posterior distribution of $Z^{(2)}$ does not depend on the health data. At the second stage, they obtain posterior samples of $f(Z^{(2)}, \theta|Y^{(1)}, Y^{(2)})$ by using $f(Z^{(2)}|Y^{(2)})$ from Stage 1 as the prior distribution of $Z^{(2)}$. Given health data $Y^{(1)}$, they assume:

$$Stage2: \quad f(Z^{(2)}, \theta|Y^{(1)}, Y^{(2)}) \propto \int f(Y^{(1)}|Z^{(2)}, \theta, \psi) f(Z^{(2)}|Y^{(2)}) f(\theta, \psi) d\psi$$

In order to decrease computational burden, they treat ψ as nuisance parameters and carry out a profile sampler approach as described in Lee et al. (2005). This approach also provides samples of $[Z^{(2)}|Y^{(1)}, Y^{(2)}]$, the posterior distribution of the average pollution exposure incorporating the health information. When the exposure model is complicated or when one is interested in running multiple epidemiology model with different sets of covariates either for a

single outcome or multiple outcomes, this two-stage approach has the advantage that one does not have to refit the exposure model when running multiple health effect analysis.

4.3 Multiple imputation

One approach to performing a two–stage analysis is to use multiple imputation (Little and Rubin, 1987). Commonly it is used to repeat data analysis on a number of datasets that contain some measure of uncertainty which arises as a result of using predictions from a model in place of missing values. It provides a method for combining the results of the repeated analyses into a summary measure with an estimate of its uncertainty. Here, at the first stage the exposure model is used to predict values at locations in space and time for which they are required by at which they may not be available. A set of samples from the posterior distributions of these predictions will be drawn to create a dataset. This is repeated to create a set of datasets, each of which will differ, representing the uncertainty associated with using predictions from the model. Rubin (1987) presented a method for combining results from the multiple analyses. Repeatedly running the health model using the different exposure datasets results in an estimate of the relative risk, β_1, and associated standard error for each dataset. These are then combined to give an overall estimate of relative risk together with a combined standard error that can be used to calculate confidence intervals. Assume β_{1d} is the estimate obtained from data set d ($d = 1, \ldots, D$) and $\sigma_{\beta d}$ is the standard deviation associated with β_{1d}. The overall estimate is the average of the individual estimates,

$$\overline{\beta} = \frac{1}{D} \sum_{1}^{D} \beta_{1d}$$

The overall estimate of the standard error will be a function of a combination of within–imputation variance and between–imputation variance. The first of these is given as

$$\sigma^2_{w\beta} = \frac{1}{D} \sum_{1}^{D} \sigma^2_{\beta d}$$

and the between-imputation variance by,

$$\sigma^2_{b\beta} = \frac{1}{D-1} \sum_{1}^{D} (\beta_{1d} - \overline{\beta}_{1d})^2$$

Then the total variance is

$$\tau^2 = \sigma^2_{w\beta} + (1 + \frac{1}{D})\sigma^2_{b\beta}$$

Confidence intervals are obtained using quantiles of the t-distribution with degrees of freedom

$$df \;=\; (D-1)\left(1+\frac{D\sigma^2_{w\beta}}{(D+1)\sigma^2_{b\beta}}\right)^2$$

4.4 Conclusion

For the fully Bayesian framework, the 'feedback' from the health model to the exposure model is conceptually not desired. This is because the exposures might be thought to cause health effects, but the health effects are not thought to affect the exposures in the same way. In addition, jointly fitting the health and exposure models is also associated with computational considerations, especially if the exposure model deals with large amounts of data over space and time. If the exposure model was complicated or multiple candidate epidemiological models were run with different sets of covariates, a single model is not going to provide an efficient method of investigation. A two-stage approach has the advantage that the exposure model is not to be refitted when running multiple health effect analyses. This significantly reduces the computational burden. Although two stage approach separates the exposure and health components, it still allows uncertainty from the exposure modelling to be incorporated into the health model. In this thesis, the first stage exposure models are analysed in Bayesian methods, while the health models are implemented in frequentist way. We use multiple imputation based on samples from the joint distribution of the posterior distributions for predictions of the exposures.

Chapter 5

Assessing the standard approach to modelling risks

The majority of studies examining the relationship between exposure to air pollution and health regress population based mortality counts against ambient pollution levels and a set of covariates (see for example Samet et al. (2000) and Katsouyanni et al. (2001)). These data typically relate to a single geographical area such as a city with the pollution levels being measured at a number of fixed locations. Most studies use a 'standard' measure of pollution exposure which is the daily average of the measurements across all monitoring sites. However, taking such an average may be a poor estimate of exposure if observations are measured with error or where there is spatial variation in the underlying pollution field which is not captured. This may lead to biases in subsequent associations with mortality. In this chapter, we present a series of simulation studies in order to assess the effects of measurement error and spatial variation on the estimates of relative risk when applying the standard modelling approach.

This chapter is split into four sections. The first section presents the basic model set up used in the simulation studies. In the second section, we present an example of implementing the health model with the estimated exposure from standard exposure model in a Bayesian setting using MCMC. Section 5.3 outlines two issues with air pollution data that may limit the efficacy of standard model; measurement error and spatial variation. In order to understand the effects of spatial variation, we look into the properties of spatial correlation in Section 5.4. The simulation studies are presented in Section 5.5. We firstly outline the method used for generating the data for simulation, followed by an assessment of the effects of measurement error and spatial variation.

5.1 Modelling setup

The model set up follows the hierarchical structure described in Chapter 2. The observation level $Y|Z, X, \theta_1$ is assumed to arise from an underlying process which is unobservable but from which measurements can be taken, with error, at locations in space and time. Here we use the notation θ_1 to denote parameters for the observation level and θ_2 for those at the process level. Measurements may also be available for covariates, X, but for clarity we do not consider covariates in the simulation studies. The underlying process level which drives the measurements seen at the observation level is denoted by $Z|\theta_2$. Models for the parameters $\theta = (\theta_1, \theta_2)$ are given in the parameter level and may control things such as variability and the strength of any spatio–temporal relationships.

The underlying spatio–temporal process, Z, may be viewed as lying in continuous domains of time and space, $\mathcal{T} \subset \mathcal{R}^1$ and $\mathcal{S} \subset \mathcal{R}^2$ respectively, where $\mathcal{R}^1$ denotes 1 dimensional Euclidean space and $\mathcal{R}^2$ denotes 2 dimensional Euclidean space. However, even when Z is continuously monitored over time, monitors may only report results at discrete times i.e. $\mathcal{T} = \{0, 1, \ldots, N_T\}$ for some number of time points N_T. The same may be true over space, where the locations where air quality monitors can actually be placed may be restricted to a relatively small number of locations, for example on public land, leading in practice to a discrete set of locations $S \in \mathcal{S}$.

The approach developed in this paper involves models for both health counts and exposures and each of these can be framed in the context of a hierarchical model (and as we describe later, these can be combined). To avoid ambiguity between the two, we use $Y^{(1)}, X^{(1)}, Z^{(1)}, \theta^{(1)}$ for the health models and $Y^{(2)}, X^{(2)}, Z^{(2)}, \theta^{(2)}$ for the exposure models. It is noted that although the health counts, $Y^{(1)}$, can be considered to be measurements from an underlying true level with differences occurring, for example due to misclassification or data anomilies, but in the absence of any other information, here we consider them to be an accurate reflection of the truth, i.e. $Y^{(1)} = Z^{(1)}$.

The standard model estimates underlying exposure levels simply by taking the average of all measured data at a particular time. The true value of the average is denoted by $\widetilde{Z}_t^{(2)}$ which here will be:

$$\widetilde{Z}_t^{(2)} = \overline{Y}_t^{(2)} = \frac{1}{N_s} \sum_{s=1}^{N_s} Y_{st}^{(2)} \quad for \quad t = 1, \ldots, N_t \tag{5.1}$$

The advantage of using this statistic is its simplicity, being a simple average that requires

no additional modelling. It is typically related to the mortality counts using Poisson linear or additive models,

$$\begin{aligned} Y_t^{(1)} &\sim Poisson(\mu_t) \qquad for \quad t = 1, \ldots, N_t \\ ln(\mu_t) &= \beta_0 + \beta_1 \widetilde{Z}_t^{(2)} \\ \theta_1 &= (\beta_0, \beta_1) \end{aligned} \tag{5.2}$$

The parameter of primary interest is β_1 which represents the association between mortality and $10\mu gm^{-3}$ increase in air pollution and mortality (at the ecological level). Since the expected value from a Poisson distribution is μ_t, and the air pollution level is rescaled by dividing by 10 in our study, so equation 3.1 becomes:

$$\begin{aligned} RR &= \frac{exp(\beta_0 + \beta_1 \frac{A+10}{10})}{exp(\beta_0 + \beta_1 \frac{A}{10})} \\ &= exp(\beta_1) \end{aligned} \tag{5.3}$$

5.2 Bayesian implementation of standard Poisson model

Here we show an example of implementation of the model given in equation 5.2 using MCMC which will be discussed in more detail in Chapter 6. For the purposes of this example, we show the full conditionals for a model without the covariate effects, i.e

$$\begin{aligned} Y_t^{(1)} &\sim Poisson(\mu_t) \qquad for \quad t = 1, \ldots, N_t \\ ln(\mu_t) &= \beta_0 + \beta_1 \widetilde{Z}_t^{(2)} \\ \beta_1 &\sim N(0, \sigma^2_{\beta_1}) \\ \beta_0 &\sim N(0, \sigma^2_{\beta_0}) \end{aligned}$$

where $\widetilde{Z}_t^{(2)}$ denotes the summary measure of the true values of the pollution field. $\sigma^2_{\beta_0}$ and $\sigma^2_{\beta_1}$ are assigned very large values, for example 10^3, so that the priors of β_0 and β_1 are vague. Since $Y_t^{(1)}$ follows a Poisson distribution with parameter μ_t, and $ln(\mu_t)$ is expressed as a function of $\widetilde{Z}_t^{(2)}$, the probability mass function of $Y_t^{(1)}$ can be written as:

$$f(Y_t^{(1)}|.) = \frac{exp(\beta_0 + \beta_1 \widetilde{Z}_t^{(2)})^{Y_t^{(1)}} exp(-exp(\beta_0 + \beta_1 \widetilde{Z}_t^{(2)}))}{Y_t^{(1)}!}$$

Then the joint distribution of (β_0, β_1) can be expressed as:

$$f(\beta_0, \beta_1 | Y_t^{(1)}, \widetilde{Z}_t^{(2)}) \quad \propto \quad \prod_{t=1}^{N_t} f(Y_t^{(1)}, \widetilde{Z}_t^{(2)} | \beta_0, \beta_1) \times f(\beta_0) \times f(\beta_1)$$

where $f(\beta_0)$ and $f(\beta_1)$ denote the prior of β_0 and β_1.

5.2.1 Sampling from the full conditional distribution of intercept term in health model, β_0

The prior for the intercept term β_0 in health model is a normal distribution,

$$f(\beta_0) \sim N(0, \sigma^2_{\beta_0}) \propto exp\Big\{ - \frac{\beta_0^2}{2 \times \sigma^2_{\beta_0}} \Big\}$$

The full conditional of β_0 is a product of all the Poisson observations and this Gaussian prior. According to Bayes' theorem, the full conditional distribution of β_0 can be written as

$$f(\beta_0 | Y_t^{(1)}, \widetilde{Z}_t^{(2)}, \beta_1) \propto \prod_{t=1}^{N_t} f(Y_t^{(1)} | \beta_0, \beta_1, \widetilde{Z}_t^{(2)}) \times f(\beta_0)$$

It follows that

$$f(\beta_0 | .) \propto \prod_{t=1}^{N_t} exp(\beta_0)^{Y_t^{(1)}} exp(- \sum_{t=1}^{N_t} exp(\beta_0 + \beta_1 \widetilde{Z}_t^{(2)})) exp\Big\{ - \frac{\beta_0^2}{2 \times \sigma^2_{\beta_0}} \Big\}$$

The Gaussian prior is not conjugate to the Poisson data, which results in a non-standard full conditional distribution. A Metropolis–Hastings algorithm is typically used for updating β_0, and a common proposal distribution is based on random walks.

5.2.2 Sampling from the full conditional distribution of relative risk, β_1

The prior for the relative risk parameter β_1 is assumed to be a vague normal prior distribution:

$$f(\beta_1) \sim N(0, \sigma^2_{\beta_1}) \propto exp\Big\{ - \frac{\beta_1^2}{2 \times \sigma^2_{\beta_1}} \Big\}$$

The posterior distribution of relative risk β_1 combines the information from the probability mass function of observed mortality and the prior of β_1. In this case, the full conditional distribution of β_1 can be written as

$$f(\beta_1|Y_t^{(1)}, \widetilde{Z}_t^{(2)}, \beta_0) \propto \prod_{t=1}^{N_t} f(Y_t^{(1)}|\beta_0, \beta_1, \widetilde{Z}_t^{(2)}) \times f(\beta_1)$$

Hence in detail, it can be written as

$$f(\beta_1|.) \propto \prod_{t=1}^{N_t} exp(\beta_1 \widetilde{Z}_t^{(2)})^{Y_t^{(1)}} exp(-\sum_{t=1}^{N_t} exp(\beta_0 + \beta_1 \widetilde{Z}_t^{(2)})) exp\Big\{ -\frac{\beta_1^2}{2 \times \sigma_{\beta_1}^2} \Big\}$$

Due to the complexity of this full conditional distribution, it is not a closed form of any familiar distribution and the Metropolis-Hastings algorithm is used to update the value of β_1 in each iteration.

5.2.3 Example

We applied the standard model to data from Greater London. This comprises daily counts of mortality for 3 years with daily measurements of PM_{10} from a set of 23 monitoring sites throughout the city. Further details can be seen in Section 8.1. The MCMC algorithm was run for 40,000 iterations discarding the first 10,000 as 'burn in'. The main interest is the estimation of the RR$= exp(\beta_1)$. Figure 5-1 shows that the samples drawn from the posterior distribution of β_1 are relatively stable and appear to converge. The mixing rate is controlled by choosing a relatively small standard deviation for the proposal distribution in the Metropolis Hastings steps. Since β_1 is log scaled relative risk in this model, the estimated relative risk is $exp(0.0293)$, that is 1.0298.

As expected, this results in the same estimate for β_1 as a standard Poisson GLM. Often a number of candidate health models may be assessed within a epidemiological analysis and the computational aspects of running a MCMC will be much greater than a standard GLM. We will develop approaches which integrate complex Bayesian exposure models with health models using standard GLMs; details of how this is achieved are given in Section 4.3

5.3 Sources of variation

We will consider limitations of using the standard approach to summarise exposures for use in health analyses with the presence of measurement error and spatial variation:

(a) - Spatial variation - If the underlying surface exhibits substantial spatial variation, the measurements at the monitor locations are unlikely to be a representative sample of the pollu-

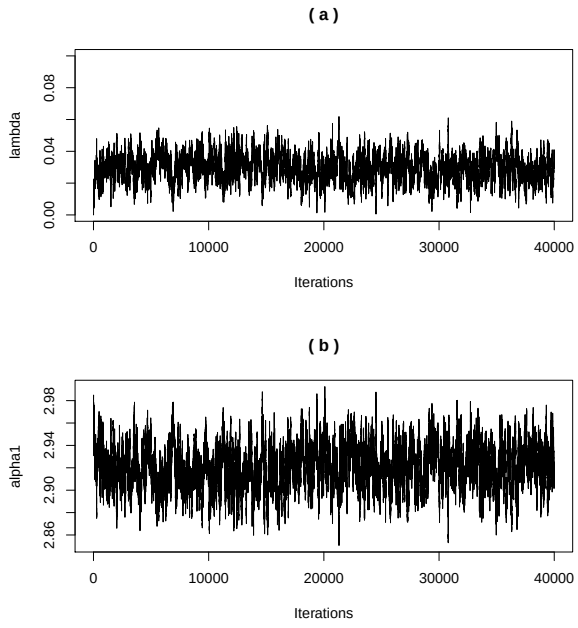

Figure 5-1: The trace plot of sampled log relative risk β_1 and β_0 using an MCMC algorithm with 40,000 iterations

tion levels throughout $\mathcal{S}$. This is because the monitor locations are likely to be small in number, unequally spaced throughout the region and located for specific reasons (for example at a well known pollution hot spot), meaning that averaging the values at these sites may not produce a good estimate of the exposure experienced by the population under study.

(b) - Measurement error - The ambient monitors are also known to measure with error (Department of the Environment, Transport and the regions 1998), meaning that $Y_{st}^{(2)}$ may be a biased estimate of the true pollution level at spatial location s on day t.

A number of authors (Duddek et al., 1995; Carlin et al., 1999; Zhu et al., 2003; Lee and Shaddick, 2010; Peng and Bell, 2010; Szpiro et al., 2011) have attempted to address these problems by replacing the average of observations with an alternative estimate, that incorporates spatial variation and measurement error in the estimate of exposure to air pollution by modelling the underlying pollution surface with a spatio-temporal model. However such models are computationally intensive to implement. In this chapter we investigate the possible effects of both these issues, investigating firstly the amount of bias each inflicts on the standard model and secondly determining whether a more complex model that, in theory, allows for these issues produces less biased results. These questions are addressed in this chapter using

simulation studies.

5.4 Spatial correlation exploration

The spatial correlation between monitoring sites will determine the nature of the pollution field. Four examples of simulated air pollution measurements from 5 monitors with different correlation–distance parameter (ϕ) in a $20km \times 20km$ region are shown in Figures 5-3 and 5-4. The spatial correlation matrix adopted in this simulation has the form of an exponential covariance function

$$\sigma_m^2 \Sigma(\phi) \; = \; \sigma_m^2 exp(-\phi \parallel d \parallel)$$

where d denotes the distance between monitoring sites, and the correlation is of the form $exp(-\phi \parallel d \parallel)$. Therefore small value of ϕ leads to high spatial correlation, with larger ϕ corresponding to lower spatial correlation.

Figure 5-2 gives an indication of the spatial correlation for different value of ϕ; from 0.01 to 0.8. It shows the effect of the parameter controlling the decay in correlation with distance. For example, the spatial correlation drops down to 0 dramatically when ϕ is large (0.8). The correlation at a selection of distances for different values of the correlation–distance parameter, ϕ, are given in Table 5.1. At a distance of 5km, the spatial correlation for ϕ equal to 0.1 and 0.01 are 0.606 and 0.951 respectively. In extreme cases of $\phi = 0.001$ and $\phi = 0.8$, the correlation will be 0.995 and 0.018 respectively. Furthermore, if the distance is extended to 15km, then the spatial correlation is 0.223, 0.861, 0.985 and 6.14×10^{-6} for ϕ equals 0.1, 0.01, 0.001 and 0.8 respectively. Figure 5-3 shows the correlations between measurements over time with data series from monitors where there is high spatial correlation being much more similar than those where there is lower spatial auto–correlation. Plots of the spatial variability are given in Figure 5-4.

5.5 Simulation studies

The effects of measurement error and spatial variation on the estimated relative risk obtained by the standard model are now investigated in a series of simulation studies. As the true risk and the characteristics of the pollution data are known, we can assess the effects of variability and error on estimation of risk. As the relative risk in real life is likely to be small, we may select a relatively large value of the true relative risk making biases easier to detect.

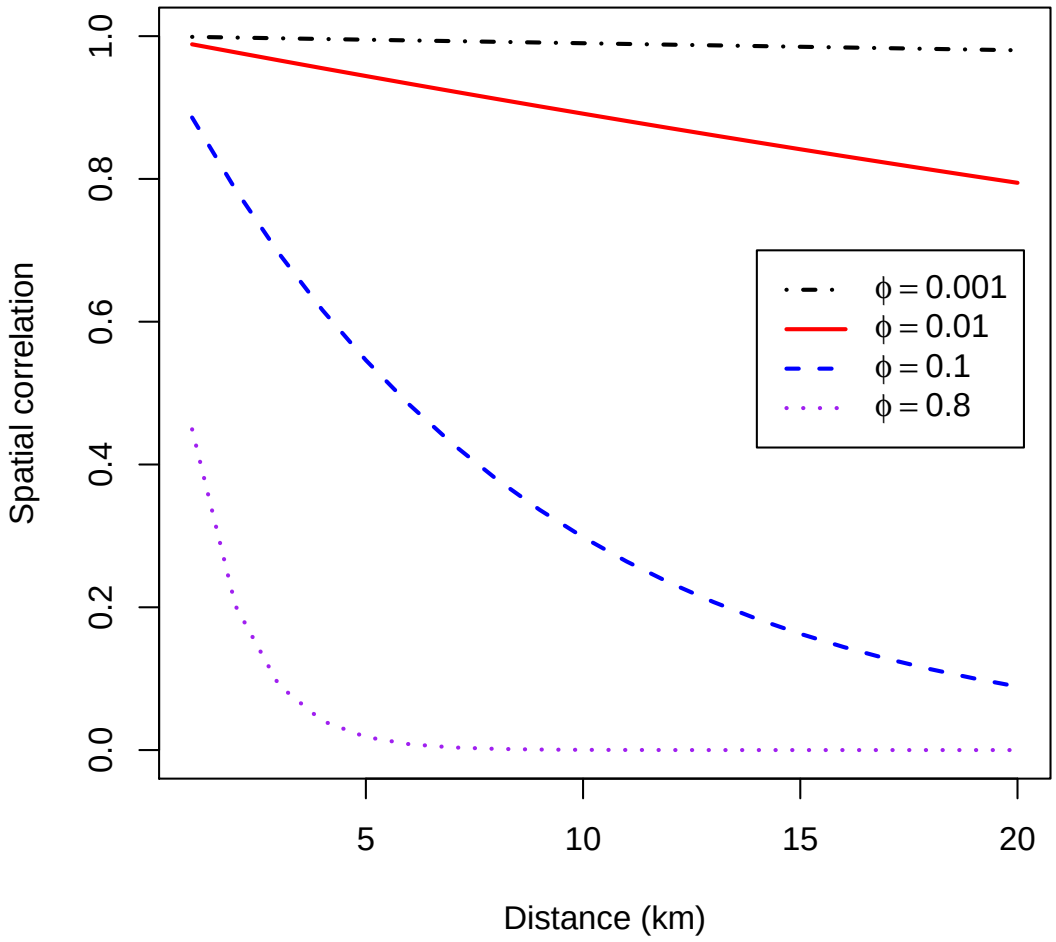

Figure 5-2: The comparison between four sets of spatial correlations with the distance from 0 to 20km by taking four different correlation–distance parameters 0.001, 0.01, 0.1 and 0.8.

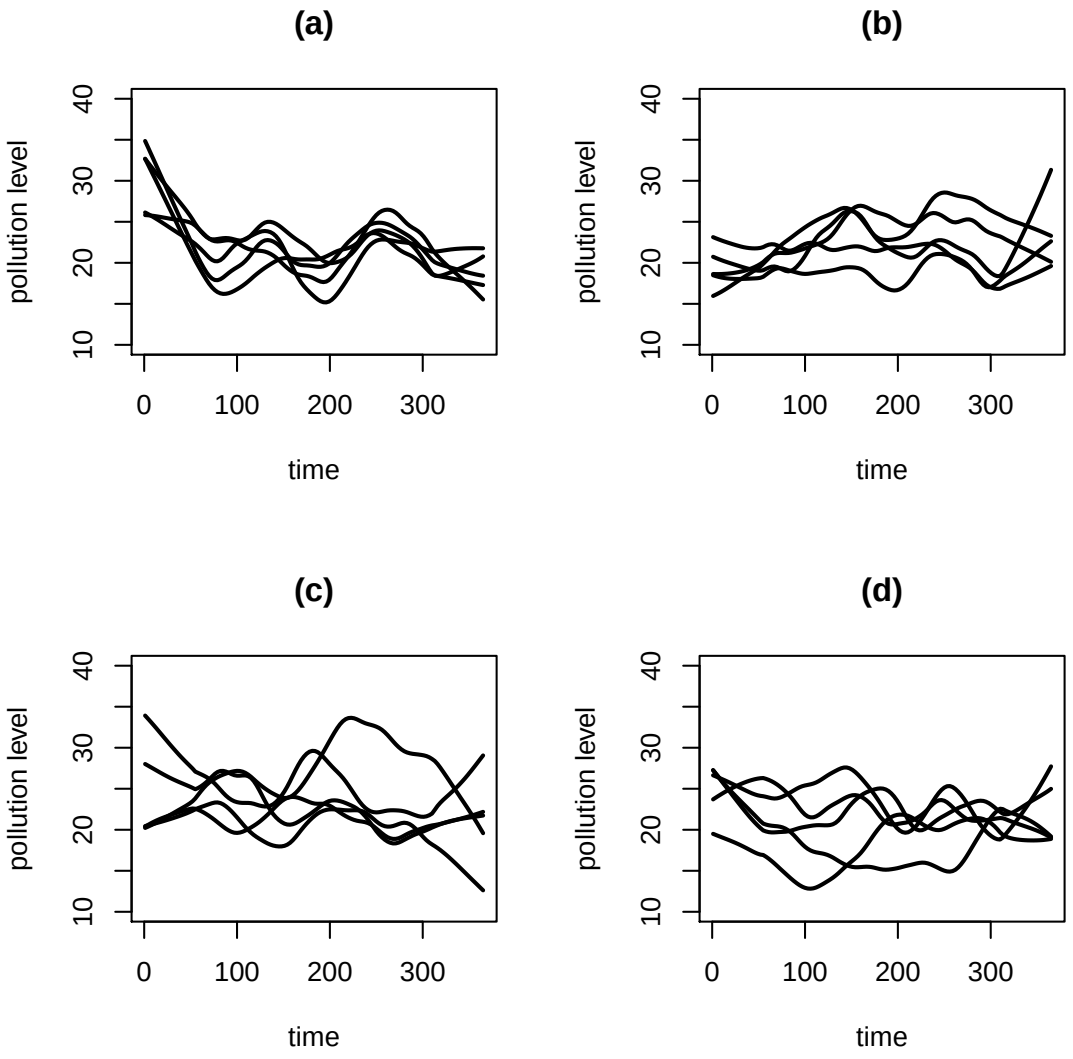

Figure 5-3: Four sets of simulated air pollution data with spatial correlation of 0.99, 0.8, 0.1 and 0.001 from 5 monitoring sites during a study time of 365 days. (a): Spatial correlation of 0.99; (b): Spatial correlation of 0.8; (c): Spatial correlation of 0.1; (d): Spatial correlation of 0.01

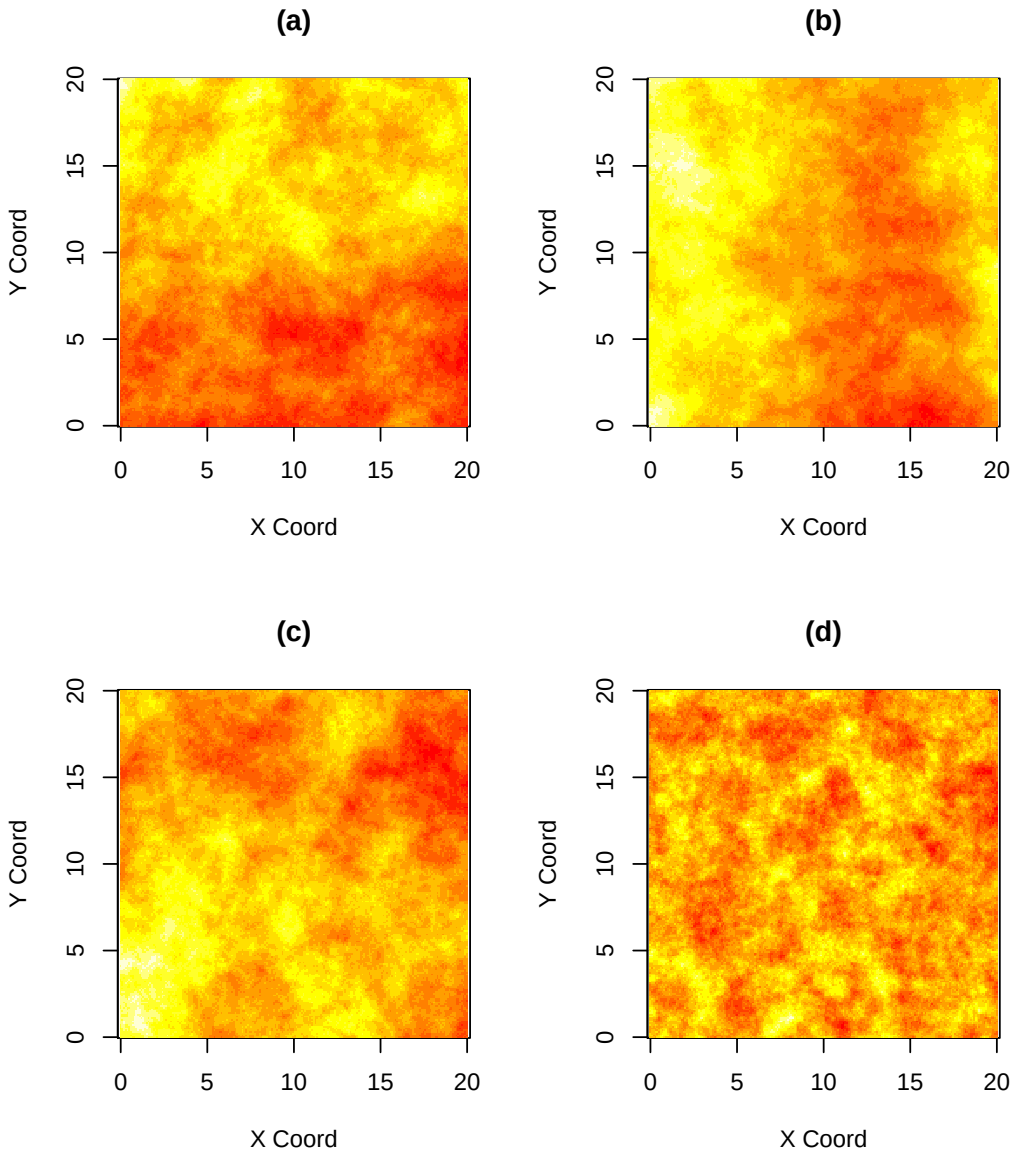

Figure 5-4: Four geographical plots of simulated spatial correlations of 0.99, 0.8, 0.1, 0.01 in a $20km \times 20km$ study region. (a): Spatial correlation of 0.99; (b): Spatial correlation of 0.8; (c): Spatial correlation of 0.1; (d): Spatial correlation of 0.01

Table 5.1: Correlation at a selection of distances for different values of the correlation–distance parameter, ϕ.

Correlation–distance parameter ϕ	Distance			
	5km	10km	15km	20km
0.001	0.995	0.990	0.985	0.980
0.01	0.951	0.905	0.861	0.819
0.1	0.606	0.368	0.223	0.135
0.8	0.018	3.35×10^{-4}	6.14×10^{-6}	1.12×10^{-7}

5.5.1 Simulation procedure

The data are generated for 365 consecutive days (1 year) on a $20km \times 20km$ square grid of 400 locations (co-ordinates (0 , 0) to (20 , 20)) at intervals of one kilometer. Two monitor placement schemes are adopted in this simulation, the first with five monitors and the second with twenty. All twenty locations are selected at random and then five are selected from them, in order to eliminate the unnecessary uncertainty that may be introduced by different location choices in different simulated datasets. The locations of the monitoring sites used in this simulation are shown in Figure 5-5.

The effects of measurement error and spatial variation on the estimated relative risks using the standard model are investigated by comparing the estimated relative risks with those known to be true. The procedure for the simulation study is as follows:

- Step 1: Generate a set of underlying levels of air pollution Z over time. This step follows the equation given below:

  $$Z_t^{(2)} \sim N(\mu + \rho(Z_{t-1}^{(2)} - \mu)\ ,\ \sigma_z^2)$$

  The exposures are assumed to have a mean level of 40 without any long-term trend, meaning that an intercept term ($\mu = ln(40/10) = 3.8$) is assigned to the temporal process. The temporal variation is induced by a first order autoregressive process specified by $\rho = 0.8$ and $\sigma_z = 0.25$. Then values are chosen to be similar to those observed in the real data used in the case study presented in Chapter 8.

- Step 2: Based on the true relative risk, $exp(\beta_1)$, generate a set of Poisson health counts, $Y_t^{(1)}$;

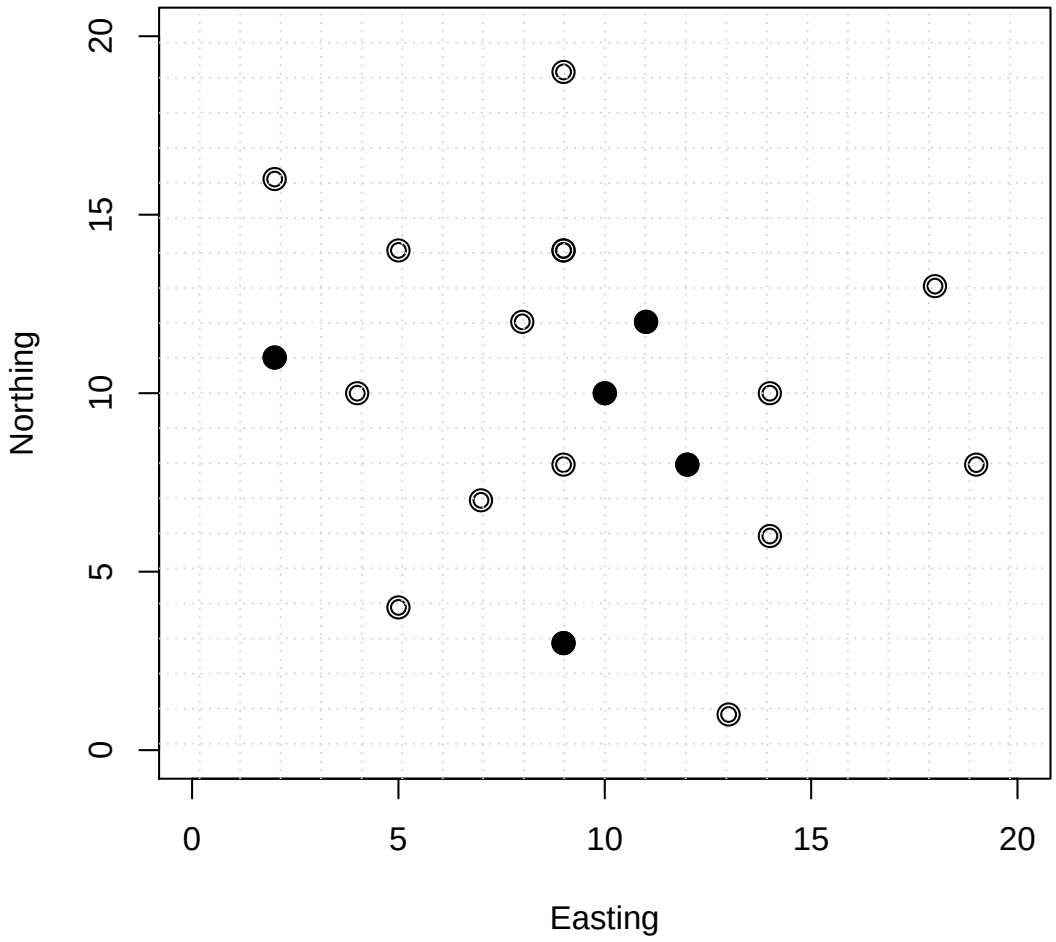

Figure 5-5: The locations of the ambient monitors in the 20km×20km grid used in this simulation study. The filled circles represent the subset of five monitors used in certain scenarios

$$\begin{aligned} Y_t^{(1)} &\sim Poisson(\mu_t) \\ ln(\mu_t) &= \beta_1 exp(Z_t^{(2)}) \end{aligned}$$

For simplicity and in order to identify the drivers of changes in the estimate of risk, the mortality data are generated without an intercept term. The association between ambient pollution levels and mortality is chosen to represent a RR of 1.2 for an increase in ten units of air pollution (that is $\beta_1 = ln(1.2)$). This is the true 'relative risk'.

- Step 3: Generate sets of measurements of air pollution using a measurement error or spatio-temporal model, the later of which will incorporate spatial variation in the underlying levels of pollution.

  The amount of measurement error in the generated air pollution data is controlled by the measurement error standard deviation σ_ϵ. In order to incorporate the measurement error effects on air pollution data, we firstly choose a sequence of values of σ_ϵ, then for each of these values, 200 set of simulated data are generated based on the equation:

$$ln(Y_{st}^{(2)}) \sim N(Z_t^{(2)}, \sigma_\epsilon^2)$$

  A sequence of values of σ_ϵ is chosen from 0 up to 2.5 at intervals of 0.1, this results in 25 different measurement error scenarios.

  The amount of spatial variation in the pollution data is controlled by the spatial standard deviation, σ_m, and correlation–distance parameter ϕ. In order to incorporate spatial variation into the air pollution data, we firstly choose a set of values of σ_m and ϕ respectively. For the choices of spatial standard deviation σ_m, we suggest a sequence from 0 up to 1.5 at intervals of 0.1. The correlation–distance parameter, ϕ, is chosen in the range 0.0115 to 0.121 as the largest distance between two participated monitoring sites locations from Figure 5-5 is about $19km$, this means that the spatial correlation will be between about 0.1 to 0.8 in the study region. For each combination of σ_m and ϕ, 200 sets of spatial effects are generated and added to the exposure data as follows:

$$
\begin{aligned}
ln(Y_{st}^{(2)}) &= Z_t^{(2)} + m_s \\
m_s &\sim MVN(0, \sigma_m^2 \Sigma_\phi)
\end{aligned}
$$

where MVN denotes the multi-variate normal distribution with covariance matrix $\sigma_m^2 \Sigma_\phi$.

- Step 4: Estimate the RR using the standard model. This averages the data generated in step 3 over space, and then uses this daily average in the health model.

$$
\begin{aligned}
Y_t^{(1)} &\sim Poisson(\mu_t) \\
ln(\mu_t) &= \beta_1 \widetilde{Z}_t^{(2)} \\
\widetilde{Z}_t^{(2)} &= \overline{Y}_t^{(2)} = \frac{1}{N_s} \sum_{s=1}^{N_s} Y_{st}^{(2)}
\end{aligned}
$$

5.5.2 The effect of measurement error

For each of the measurement error scenarios, the relative risks are estimated using the 200 data sets using the standard model. Average exposures are used based on the measurements from both five and twenty monitors. The estimated relative risks are shown in Figure 5-6 and Table 5.2. For Figure 5-6, the black line represents the median (over 200 data sets) for each level of measurement error. The dark and light shading indicate the middle 50% and 95% of the distribution of RRs respectively.

Table 5.2: Summary of the estimated relative risks from data with measurement error based on 5 or 20 monitoring sites: medians from values from 200 simulated datasets.

No. of monitors	Measurement error standard deviations					
	0.3	0.5	1.0	1.5	2.0	2.3
5	1.1978	1.1831	1.1305	1.0801	1.0454	1.0311
20	1.2043	1.2005	1.1834	1.1581	1.1304	1.1124

Figure 5-6 shows the impact that measurement error can have on the the relationship between air pollution measurements and mortality. Attenuation to no effect can clearly be seen as increasing levels of measurement error level lead to the shrinking of the estimated relative

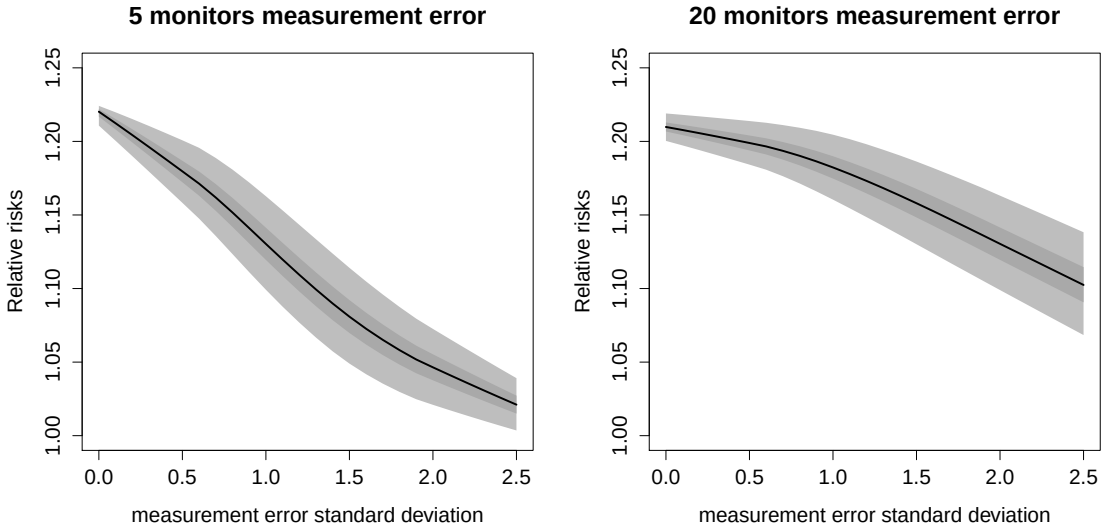

Figure 5-6: The estimated relative risks summarized form 200 simulations for each value of measurement error standard deviation. (a) 5 monitors (b) 20 monitors. The solid black line represents the median, whilst the grey shading covers 95% quantiles, and dark grey shading area represents 50% quantiles.

risk towards one (Carroll et al., 2010). From Table 5.2, we can see the median of the estimated relative risk starts from 1.2043 when the measurement error standard deviation is 0.3 for the 20 monitoring sites scenario, and it goes down to 1.1124 when measurement error standard deviation increases to 2.3. Similarly, the median of the estimated relative risk of 5 monitoring is 1.1978 for measurement error standard deviation equal 0.3, then it decreases massively to 1.0311 when the measurement error standard deviation is 2.3. Therefore, if measurement error increases to a certain level, then there may be no traceable connection between response and exposures, ie RR becomes close to 1.

Figure 5-6 and Table 5.2 also show the impact of number of monitors on the estimation of relative risks. The median of estimated relative risk from 20 monitoring sites scheme is closer to the true value of 1.2 than is the one estimated from 5 monitors. The mean of the estimated relative risk decays to one more quickly if five monitors are used compared with twenty. For example, for $\sigma_\epsilon = 0.3, 0.5, 1$ the mean relative risks for five monitors are 1.1978, 1.1831 and 1.1305 respectively, and 1.2043, 1.2005 and 1.1834 for the twenty monitors scheme. Although the exposures over all locations are firstly averaged and then plugged into the health model, the uncertainty caused by the measurement error is reduced by the additional information from the extra monitoring sites.

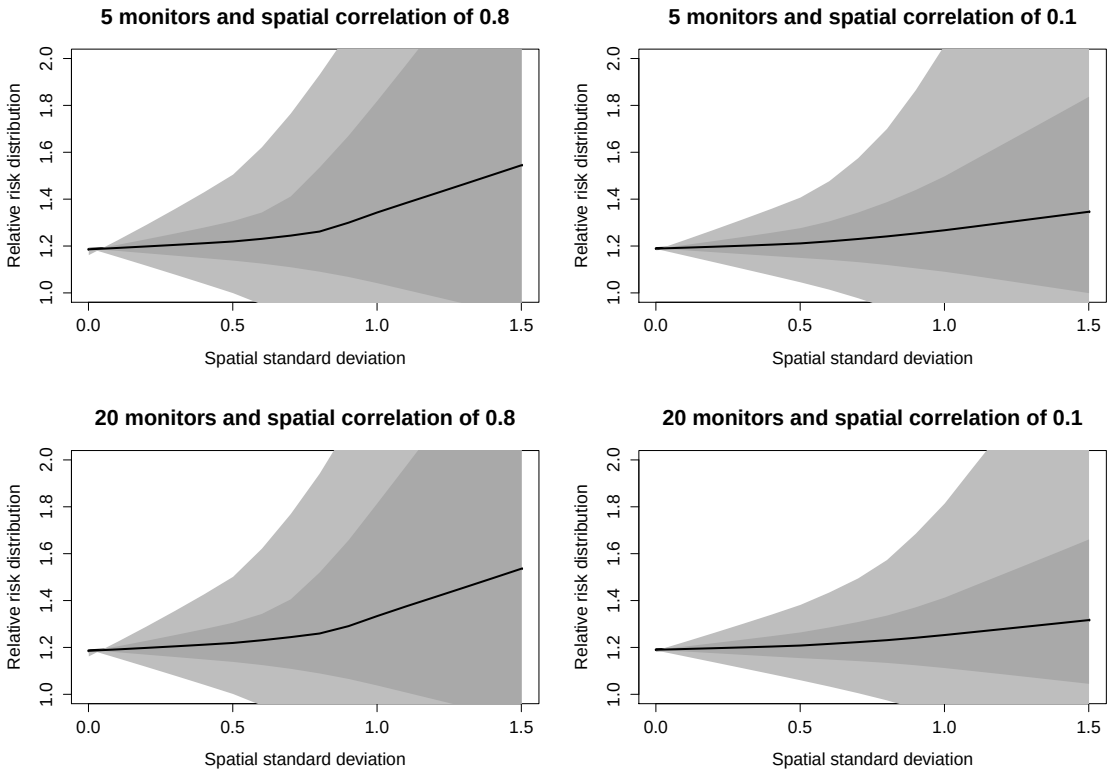

Figure 5-7: Distribution (over 200 simulated data sets) of relative risks for various levels of spatial variation and correlation–distance parameters. The solid black line represents the median, and the grey shading illustrates the variability over the 200 simulations. (a) 5 monitors and spatial correlation of 0.8; (b) 5 monitors and spatial correlation of 0.1; (c) 20 monitors and spatial correlation of 0.8; (d) 20 monitors and spatial correlation of 0.1.

5.5.3 The effect of spatial variation

For the spatial variation scenarios, relative risks are estimated from the 200 data sets using the standard model for both the five and twenty monitor cases. The estimated relative risks are given in Figure 5-7 and Table 5.3. In Figure 5-6, the black line represents the median RR (over 200 data sets) for each value of the spatial standard deviation and correlation–distance parameter ϕ. The dark and light shading cover the middle 50% and 95% of the distribution of RRs respectively.

Figure 5-7 shows that the medians of estimated relative risks in each of the four plots are very close to the true value 1.2 if there is no spatial variation, and an increase in spatial variation leads to an increase in the estimated relative risks. For example, panel (a) shows that

the medians of estimated relative risks for 5 monitors and high spatial correlation are lifted dramatically from the origin 1.2 to 1.526 when σ_m reaches 1.5, meanwhile the median of the estimated relative risks for low spatial correlation is 1.339. In addition, for the case of 20 monitoring sites, the medians of the high and low spatial correlated scenarios are 1.499 and 1.281 respectively when σ_m takes the value of 1.5. Therefore, we can see that the medians of estimated relative risks are flatter (closer to true value 1.2) for the low spatial correlation scenario where there are more monitoring sites.

From the gray shaded area in Figure 5-7, we can see the variability of estimated relative risks over 200 simulations with high spatial correlation (0.8) is larger than that induced from low correlated datasets (0.1), hence the corresponding 95% quantiles expand much faster when spatial variation increases. For example, in the simulations with 20 monitors, the 95% quantile summarised from the high spatial correlation scenario is (1.087, 1.357) when σ_m equals 0.3, then it expands to (0.684, 2.536) when σ_m increases to 1, finally the quantile reaches (0.334, 7.485) for $\sigma_m = 1.5$. By contrast, the 95% quantile for low spatial correlation scenario is (1.120, 1.293) which is much narrower than the quantile with the same level of spatial standard deviation of 0.3, eventually it expands to (0.570, 3.504) at the end for σ_m equals 1.5, this is almost the half of the width of the one from high spatial correlation scenario.

In addition, the impact of the number of monitors on the estimated relative risks is also shown in Figure 5-7 and Table 5.3. Although the number of locations has very little effect on the relative risks estimated from high spatial correlated data, it significantly impacts on the relative risks with low spatial correlation. The 95% quantiles are narrowed when the monitor numbers increased form 5 to 20.

One might initially expect the change in the quantiles for RR in Figure 5-7 to be the opposite way, i.e that the variability of the estimated relative risk from the high spatial correlated scenario will be smaller than those where there is low spatial correlation. However high spatial correlation in the air pollution data means the simulated datasets will be more similar than those with low spatial correlation. Here the shape of shaded area represents the quantile of the estimated relative risk from 200 simulated data sets for each scenario, hence it actually demonstrates the variability of the estimated relative risks between each simulation data set. In this case, the variability within a single dataset generated using high spatial correlation may be very small, but the variability among the 200 simulated datasets may be quite large. By contrast, the low spatial correlated data may result in high variability within a single dataset, but overall the datasets are more likely to be similar, for example, the means will be more similar even though the variation around then may be greater. Therefore, the phenomenon shown in Figure

Table 5.3: Summary of the estimated relative risks from data with spatial variation based on 5 or 20 monitoring sites: medians from values from 200 simulated datasets.

Monitors	Correlation	Spatial standard deviations			
		0.3	0.5	1.0	1.5
5	0.1	1.205 (1.113, 1.304)	1.207 (1.045, 1.394)	1.260 (0.884, 1.795)	1.339 (0.462, 4.767)
	0.8	1.215 (1.086, 1.358)	1.241 (0.968, 1.589)	1.318 (0.673, 2.582)	1.526 (0.323, 7.748)
20	0.1	1.203 (1.120, 1.293)	1.206 (1.066, 1.364)	1.247 (0.924, 1.681)	1.281 (0.570, 3.504)
	0.8	1.215 (1.087, 1.357)	1.239 (0.972, 1.580)	1.317 (0.684, 2.536)	1.499 (0.334, 7.485)

5-7 actually make senses; 95% quantiles of high spatial correlated datasets are wider than the ones from low spatial correlated scenarios.

5.5.4 The effects of spatial variation and measurement error

In this section, we examine the interaction between measurement error and spatial variation and the effects on the estimated relative risks. The estimated relative risks are given in Figure 5-8, Table 5.4 and Table 5.5. For Figure 5-8, the black line represents the median RR (over 200 data sets) for each chosen value of spatial standard deviation and correlation–distance parameter, ϕ. The dark and light shading cover the middle 50% and 95% of the distribution of RRs respectively.

Here we adopt four scenarios; the first two fix the spatial standard deviation as 0.5 with two options for the correlation–distance parameter, ϕ (representing low and high spatial correlations), while the measurement error standard deviation varies from 0 to 2.5. These two scenarios are designed to check the effect of measurement error on the estimated relative risks with a certain level of spatial variation in the data. The spatial standard deviation is chosen to be 0.5 because Figure 5-7 shows there is a significant change in relative risks when the value of the spatial standard deviation is greater than or equal to 0.5. The next two scenarios fix the measurement error standard deviation as 1 while varying the spatial standard deviation from 0 to 1.5 using the same two values of ϕ. These two scenarios are used to examine the effect of spatial variation on the estimated relative risks with a certain level of measurement error in the air pollution data. Furthermore, we only look into 20 monitoring scheme in order to focus on the effects of the interaction between measurement error and spatial variation.

Table 5.4: Summary of the estimated relative risks from data with measurement error and spatial variation based on 20 monitoring sites and fixing spatial standard deviation as 0.5: medians from values from 200 simulated datasets.

Correlation	Measurement error standard deviations			
	0.5	1.0	1.5	2
0.1	1.196 (1.063, 1.347)	1.182 (1.052, 1.328)	1.147 (1.045, 1.260)	1.133 (1.042, 1.232)
0.8	1.199 (1.001, 1.437)	1.186 (1.010, 1.392)	1.177 (0.966, 1.352)	1.137 (0.99, 1.299)

Compared to the plots in Figure 5-6, the top two plots of Figure 5-8 show there is only a slight decrease in the estimated relative risks and the width of the gray shading area reduces as the measurement error increases. Moreover, the width of the quantile is much larger than the ones in Figure 5-6 when no measurement error is present. In addition, the quantiles of the high spatial correlation scenarios are wider than the ones in the scenarios with low spatial variation. The bottom two panels are very similar to the ones shown in Figure 5-7, we expect the estimated relative risk is lower than the true value 1.2 when no spatial variation is present. Furthermore, the quantiles for the low spatial correlation scenarios are narrower than the ones in Figure 5-7.

We have shown that the effects of measurement error lead to attenuation of the relative risk towards one, whereas spatial variation induces an increase in the relative risks. In the top two panels of Figure 5-8, although the value of the spatial standard deviation is fixed at a moderate level of 0.5, the relative risks do not suffer from as much attenuation. Moreover, the width of the quantile when no measurement error is present is very similar to the one shown in Figure 5-7 when the spatial variance equals 0.5. By contrast, in the bottom two plots, there is not much sign of the effects of the measurement error, but the lower value of the estimated relative risk when there is no measurement error suggests a weaker effect of measurement error in the way it has attenuated the relative risk. There is likely to be a trade off between measurement error and spatial variation, and we have seen that for measurement error to dominate, it needs to be relatively very large, otherwise, the spatial variation appears to dominate.

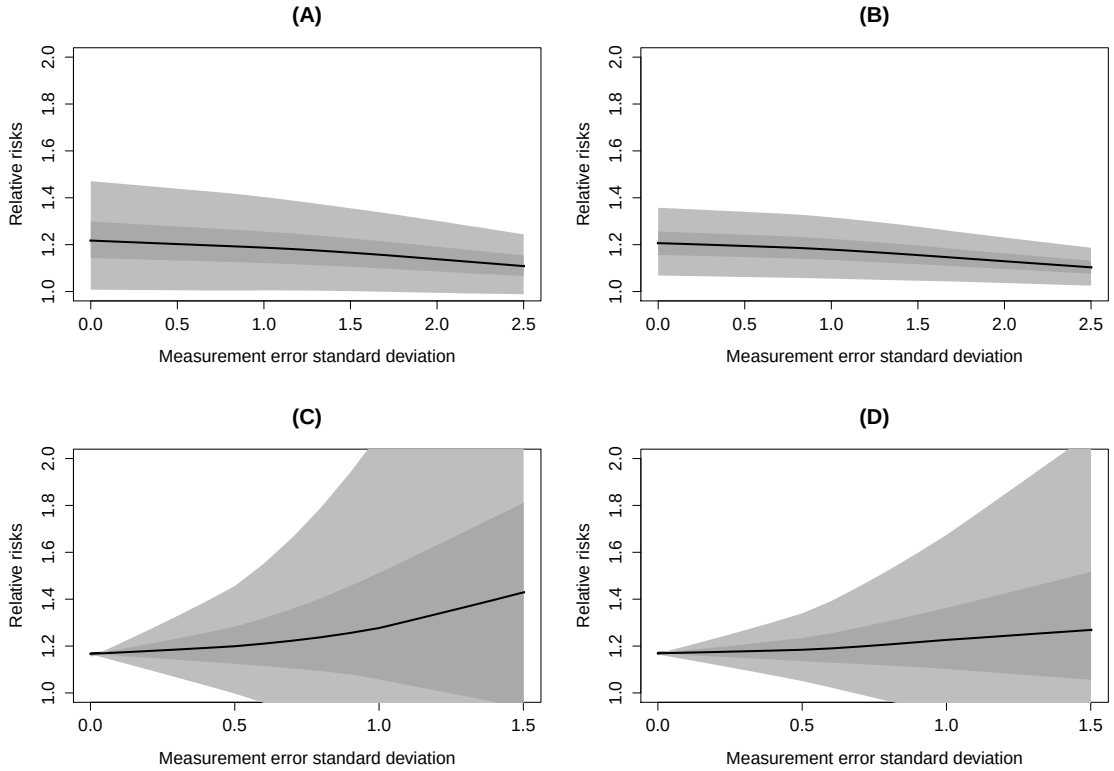

Figure 5-8: The estimated relative risks for different levels of spatial variation and measurement error. The solid black line represents the median, and the grey shading illustrates the overall 95% quantile, the dark gray area represents 50% quantile. (a) Fixing spatial standard deviation as 0.5, and correlation–distance parameter ϕ as 0.0115 (representing high spatial correlation), meanwhile, change the amount of measurement error from 0 to 2.5; (b) Fixing spatial standard deviation as 0.5, and correlation–distance parameter ϕ as 0.121 (representing low spatial correlation); (c) Fixing measurement error as 1, then change spatial standard deviation from 0 to 1.5 with ϕ equals 0.0115; (d) Fixing measurement error as 1, then change spatial standard deiviation from 0 to 1.5 with ϕ equals 0.121.

Table 5.5: Summary of the estimated relative risks from data with measurement error and spatial variation based on 20 monitoring sites and fixing measurement error standard deviation as 1: medians from values from 200 simulated datasets.

Correlation	Spatial standard deviations			
	0.3	0.5	1.0	1.5
0.1	1.177 (1.099, 1.262)	1.176 (1.061, 1.303)	1.233 (0.894, 1.701)	1.274 (0.741, 2.190)
0.8	1.186 (1.074, 1.309)	1.183 (1.001, 1.397)	1.270 (0.756, 2.132)	1.440 (0.305, 6.801)

5.6 Conclusion

In this chapter we investigated how well the standard model did in the presence of underlying variability in the air pollution data. For simulated pollution data that are spatially flat and exhibit no measurement error, taking average of measured air pollution levels over all monitoring sites is a relatively adequate representation of overall exposures, and causes little or no bias in the estimated relative risks, while in all other situations the standard method is less accurate. We found that a high level of measurement error typically attenuates the estimated relative risks towards one, whereas spatial effect inflates the values. This implies that the estimation of relative risk may be located at the true value even in the presence of both spatial effect and measurement error. Nevertheless, this is only a special case caused by the trade-off of measurement error against spatial effect, and highly depends on the value of them. Therefore, more complex structured model is suggested which allows the underlying variability to be modelled.

From the last simulation study we can see that adding spatial effect to measured variable pollution data changes the bias in relative risk significantly, whereas the addition of measurement error to data contaminated by spatial effect has little or no influence. These findings indicate that the spatial effect is more detrimental to the accuracy of the standard estimate than measurement error. The simulation studies also show that if either factor is present, the average bias drops in the estimated relative risks if twenty sets of observations are available in comparison with only five. Hence we may expect the larger number of monitors participates in the epidemiology study in practice, the more accurate results may be obtained.

The adequacy of standard model requires a trade-off of simplicity against accuracy. The differences in accuracy depend on the amount of spatial variation and measurement error in the observed data. If both factors are low then standard model is adequate, and can be used with its advantage of simplicity.

Chapter 6

Implementing Bayesian exposure models using MCMC

In this chapter, we present the details of the MCMC algorithms for the exposure models described in Chapter 6, including the form of the full conditional distributions and posterior distributions.

Markov chain Monte Carlo simulation is based on iteratively sampling from a Markov chain, $\theta^1, \theta^2, \theta^3$..., whose equilibrium distribution is the desired distribution. The constructed Markov chain is initialised by a starting value, θ_0, and is run until it has converged to its target distribution. Then the convergence can be assessed by visual observation of the trace plot and computing the Gelman and Rubin (1992) statistic. The period before convergence is called 'burn-in' and samples from this period are discarded. Although Markov chain simulation can be implemented for complex statistical models, it may suffer from poor mixing which results in high autocorrelation in the series, meaning that a sample provides rather less information about the posterior distribution than a sample of the same size containing independent observations. In addition, it may be unclear that the entire posterior distribution is explored, meaning that some possible states may be missed in the process. We now give a brief review of two methods for obtaining samples from the posterior distributions using MCMC; the Metropolis-Hastings algorithm and Gibbs sampling. Further details can be found in Gelman (2003).

6.1 Metropolis-Hastings algorithm

The Metropolis Hastings algorithm provides a procedure for exploring the posterior distribution by drawing samples from a random proposal density. The key of the Metropolis Hastings

algorithm is that a candidate is chosen with a certain probability for the next sample value based only on the current sample at each iteration. The details of the algorithm are as follows.

1. Arbitrarily draw a staring point θ^0 for the Markov chain ensuring that its posterior probability $f(\theta^0|x)$ is positive.

2. At each iteration $t-1$ (for $t = 2, 3, \ldots$), generate a candidate θ^* from a proposal distribution $f(\theta^*|\theta^{t-1})$ that is based on the current value of the Markov chain. Then the candidate is accepted with probability

$$\alpha(\theta^{t-1}, \theta^*) \; = \; min\left(1, \frac{f(\theta^*)f(\theta^{t-1}|\theta^*)}{f(\theta^{t-1})f(\theta^*|\theta^{t-1})}\right)$$

where $f(\theta^{t-1})$ represents the density function of the current state. Next, a value u is draw from the uniform distribution $U(0, 1)$. We accept the candidate as the next iteration, that is $\theta^* = \theta^t$, if $u \leq \alpha(\theta^{t-1}, \theta^*)$; otherwise, reject the proposal and the candidate value is discarded, hence the current value is reused in the next iteration, that is set $\theta^t = \theta^{t-1}$.

3. Repeat step 2 until the sequence of drawn samples reaches convergence.

MCMC can be complex to implement, and the results can be affected by the choice of starting values. Although the starting point should not affect the stationary distribution, especially for the rapidly mixing chains, it may need to be chosen carefully for slowly mixing chains which can often stick in a small area of the parameter space for a long time, thus taking much more time to reaches convergence.

The choice of proposal distribution can have a large impact on the convergence and acceptance rate of Metropolis Hastings algorithm. Specifically, the acceptance rate is significantly influenced by the variance of proposal distribution, thus a cautious proposal distribution with relatively small variance generates small steps, thereby leading to high acceptance rates. Furthermore, although a proposal distribution with large variance induces large movement, it may be trapped at the tails of the posterior distribution, thereby leading to low acceptance rates and the chain barely moving. Typically, $20 - 30\%$ of acceptance rate is considered reasonable (Gelman et al., 1995).

6.2 Gibbs sampling

The Gibbs sampler is a special case of the Metropolis Hastings algorithm, in which the proposal distribution is given by $f(\theta_k^*|\theta_k^{t-1}) = f(\theta_k^{t-1}|\theta_{-k}^{t-1})$, the full conditional distribution of θ_k (where $\theta_{-k} = (\theta_1, \ldots, \theta_{k-1}, \theta_{k+1}, \theta_n)$). In this case, the acceptance probability is simplified to 1, which effectively removes the accept or reject stage; hence movement of the Markov chain is ensured. For a vector random variable θ with joint density $f(\theta) = f(\theta_1, \theta_2, \ldots, \theta_n)$, suppose the full conditional distribution $f(\theta_k|\theta_{-k})$ for each parameter is known, then the Gibbs sampler algorithm works as follows:

1. Arbitrarily choose a starting point $\theta^{(0)} = \left(\theta_1^0, \ldots, \theta_n^0\right)$ with $f(\theta^0) > 0$

2. At step t, draw θ_k^t from the full conditional distribution $f(\theta_k|\theta_{-k}^{t-1})$

3. Then repeat step 2 until the sequence converges.

6.3 Measurement error model

We now give details of the MCMC algorithm used when fitting the models in Chapter 6. We start with the measurement error model from Section 7.4. As the following descriptions are only related to the exposure model, for clarity we drop the $Y^{(i)}$ and $Z^{(i)}$ notation we use to distinguish between health and exposure models. The measurement error exposure model can be written as:

$$\begin{aligned}
ln(Y_{st}) &\sim N(Z_t, \sigma_\epsilon^2) \quad for \;\; t = 1, \ldots, N_t, \\
Z_t &\sim N(\mu + \rho(Z_{t-1} - \mu), \sigma_z^2) \\
\sigma_\epsilon^2 &\sim Inverse - Gamma(a_\epsilon, b_\epsilon) \\
\sigma_z^2 &\sim Inverse - Gamma(a_z, b_z) \\
\rho &\sim Uniform(a_\rho, b_\rho) \\
\mu &\sim N(0, \psi^2)
\end{aligned}$$

Prior distributions for the parameters of the model are chosen as follows. The prior choice of parameter ρ is chosen as $Uniform(a_\rho, b_\rho)$, this is because ρ represents rate in a autoregressive process, so its possible range is limited between 0 and 1, therefore $a_\rho = 0$ and $b_\rho = 1$. Inverse-Gamma distribution is assigned to both σ_z^2 and σ_ϵ^2. The rest of the parameters are assigned vague priors; Gaussian distributions with large variance ψ^2. In the MCMC algorithm

for the measurement error model, the conditional probability density function of the log of outcome Y_{st} given all the other parameters is the follows:

$$f(ln(Y_{st})|.) \propto \left(\sigma_\epsilon^2\right)^{-\frac{1}{2}} exp\left\{ - \frac{[ln(Y_{st}) - Z_t]^2}{2\sigma_\epsilon^2} \right\}$$

Hence the joint posterior distribution of $(\mu, \rho, Z, Z_0, \sigma_\epsilon^2, \sigma_z^2)$ is given by

$$\begin{aligned} f(\mu, \rho, Z_t, Z_0, \sigma_\epsilon^2, \sigma_z^2 | ln(Y)) \quad \propto \quad & \prod_{s=1}^{N_s} \prod_{t=1}^{N_t} f(ln(Y_{st})|Z_t, \sigma_\epsilon^2) \times \prod_{t=1}^{N_t} f(Z_t|\mu, \rho, \sigma_z^2) \\ & \times f(Z_0|\mu, \rho, \sigma_z^2) \times f(\mu) \times f(\rho) \times f(\sigma_\epsilon^2) \times f(\sigma_z^2) \end{aligned}$$

where $f(\mu)$, $f(\rho)$, $f(\sigma_\epsilon^2)$ and $f(\sigma_z^2)$ are the priors for the uncertainties and parameters.

6.3.1 Sampling from the full conditional distribution of underlying value, Z_t

The values of the underlying level, Z_t, have the distribution function:

$$f(Z|\mu, \rho, \sigma_z^2) = \prod_{t=1}^{N_t} f(Z_t|Z_{t-1}, \mu, \rho, \sigma_z^2)$$

Each of the terms in the product can be written as

$$f(Z_t|\mu, \rho, \sigma_z^2) = \frac{1}{\sqrt{2\pi\sigma_z^2}} exp\left\{ - \frac{[Z_t - \mu - \rho(Z_{t-1} - \mu)]^2}{2\sigma_z^2} \right\}$$

The information about Z_0 is required in order to obtain the full conditional distributions. Therefore, a prior is assigned to it such as:

$$f(Z_0|\mu, \rho, \sigma_z^2) = \frac{\sqrt{(1-\rho^2)}}{\sqrt{2\pi\sigma_z^2}} exp\left\{ - \frac{(Z_0 - \mu)^2(1-\rho^2)}{2\sigma_z^2} \right\}$$

Hence the full conditional distribution of Z_t for $t = 1, \ldots, N_t - 1$ can be written as

$$\begin{aligned} f(Z_t|ln(Y_{st}), \mu, \rho, Z_{t-1}, Z_{t+1}, \sigma_z^2) \quad \propto \quad & \prod_{s=1}^{N_s} f(ln(Y_{st})|Z_t, \sigma_\epsilon^2) \times f(Z_t|\mu, \rho, Z_{t-1}, \sigma_z^2) \\ & \times f(Z_{t+1}|\mu, \rho, Z_t, \sigma_z^2) \end{aligned}$$

which will be of the form:

$$f(Z_t|.) \propto exp\Bigg\{ - \frac{Z_t^2 - 2\Big\{ \frac{\sigma_z^2 \sum_{s=1}^{N_s} ln(Y_{st}) + \sigma_\epsilon^2[\rho(Z_{t-1}+Z_{t+1}) + \mu(1-\rho)^2]}{N_s\sigma_z^2 + (1+\rho^2)\sigma_\epsilon^2 + \sigma_z^2\sigma_\epsilon^2} \Big\} Z_t}{\frac{2 \times \sigma_z^2\sigma_\epsilon^2}{N_s\sigma_z^2 + (1+\rho^2)\sigma_\epsilon^2 + \sigma_z^2\sigma_\epsilon^2}} \Bigg\}$$

This is a closed form of the normal distribution, hence each Z_t can be updated using Gibbs sampler from the posterior distribution

$$Z_t|. \sim \quad N\Bigg(\quad \frac{\sigma_z^2 \sum_{s=1}^{N_s} ln(Y_{st}) + \sigma_\epsilon^2[\rho(Z_{t-1} + Z_{t+1}) + \mu(1-\rho)^2]}{N_s\sigma_z^2 + (1+\rho^2)\sigma_\epsilon^2 + \sigma_z^2\sigma_\epsilon^2}, \\ \frac{\sigma_z^2\sigma_\epsilon^2}{N_s\sigma_z^2 + (1+\rho^2)\sigma_\epsilon^2 + \sigma_z^2\sigma_\epsilon^2} \Bigg)$$

The special cases are the posterior distributions of Z_0 and Z_{N_t}, the former of which borrows information from Z_1, while the latter only involves the value before it which is Z_{N_t-1}. Hence they can be written as:

$$Z_{N_t}|. \sim N\Bigg(\frac{\sigma_z^2 \sum_{s=1}^{N_s} ln(Y_{sN_t}^{(2)}) + \sigma_\epsilon^2[\rho Z_{N_t-1} + \mu(1-\rho)]}{N_s\sigma_z^2 + \sigma_\epsilon^2 + \sigma_z^2\sigma_\epsilon^2}, \frac{\sigma_z^2\sigma_\epsilon^2}{N_s\sigma_z^2 + \sigma_\epsilon^2 + \sigma_z^2\sigma_\epsilon^2} \Bigg)$$

$$Z_0|. \sim N\Big(\mu + \rho(Z_1 - \mu), \sigma_z^2\Big)$$

6.3.2 Sampling from the full conditional distribution of the autoregressive process intercept, μ

The prior for the intercept term μ is assumed to be vague

$$\mu \sim N(0, \psi^2) \propto exp\Big\{ - \frac{\mu^2}{2 \times \psi^2} \Big\}$$

The full conditional distribution of μ only borrows information from the autoregressive process, hence it can be written as:

$$f(\mu|Z, \rho, \sigma_z^2) \sim \prod_{t=1}^{N_t} f(Z_t|Z_{t-1}, \mu, \rho, \sigma_z^2) \times f(Z_0|\mu, \rho, \sigma_z^2) \times f(\mu)$$

As a result the posterior distribution of μ is:

$$\mu|. \sim N\left(\frac{\psi^2[(1-\rho)\sum_{t=1}^{N_t}(Z_t-\rho Z_{t-1})+(1-\rho^2)Z_0]}{\sigma_z^2+\psi^2(1-\rho^2)+\psi^2 N_t(1-\rho)^2}, \frac{\psi^2\sigma_z^2}{\sigma_z^2+\psi^2(1-\rho^2)+\psi^2 N_t(1-\rho)^2}\right)$$

6.3.3 Sampling from the full conditional distribution of the autoregressive process parameter, ρ

The autoregressive process parameter ρ is assigned a continuous uniform prior

$$\rho \sim Uniform(a_\rho, b_\rho)$$

Since the information on Z_0 is involved in the full conditional distribution of ρ, we can not obtain the posterior distribution directly. In this case, the Metropolis Hastings algorithm is adopted to generate samples of ρ. In order to simplify the process, the proposal distribution is chosen as:

$$\rho|Z,\mu,\sigma_z^2,Z_0 \sim N\left(\frac{\sum_{t=1}^{N_t}(Z_{t-1}-\mu)(Z_t-\mu)}{\sum_{t=1}^{N_t}(Z_{t-1}-\mu)^2}, \frac{\sigma_z^2}{\sum_{t=1}^{N_t}(Z_{t-1}-\mu)^2}\right)$$

In this circumstance, the acceptance ratio only depends on the density function of Z_0, therefore, the proposed value of ρ is accepted with probability:

$$c = min\left\{1, \frac{f(Z_0|\rho^\star,\mu,\sigma_z^2)}{f(Z_0|\rho^c,\mu,\sigma_z^2)}\right\}$$

where $\rho^\star$ denotes the proposed value of ρ, and ρ^c represents the current value of ρ.

6.3.4 Sampling from the full conditional distribution of variance of measurement error, σ_ϵ^2

The hyper-prior for the variance of measurement error σ_ϵ^2 is an Inverse-Gamma distribution

$$\sigma_\epsilon^2 \sim Inverse-Gamma(a_\epsilon, b_\epsilon)$$

where a_ϵ and b_ϵ are chosen to be very small. The posterior distribution is:

$$f(\sigma_\epsilon^2|ln(Y),Z) \propto \prod_{t=1}^{N_t}\prod_{s=1}^{N_s} f(ln(Y_{st})|Z_t,\sigma_\epsilon^2) \times f(\sigma_\epsilon^2)$$

It follows the form

$$f(\sigma_\epsilon^2|.) \propto (\sigma_\epsilon^2)^{-a_\epsilon - \frac{N_s N_t}{2}} exp\left\{ - \frac{\sum_{t=1}^{N_t} \left[\sum_{s=1}^{N_s} (ln(Y_{st}) - Z_t)^2 \right] + 2b_\epsilon}{2\sigma_\epsilon^2} \right\}$$

This is a kernel of an Inverse-Gamma distribution:

$$\sigma_\epsilon^2|. \sim Inv - Gam\left(a_\epsilon + \frac{N_s N_t}{2}, b_\epsilon + \frac{\sum_{t=1}^{N_t} \left[\sum_{l=1}^{N_s} (ln(Y_{st}) - Z_t)^2 \right]}{2} \right)$$

6.3.5 Sampling from the full conditional distribution of the variance of the autoregressive process, σ_z^2

The hyper-prior for the variance of measurement error σ_z^2 is an Inverse-Gamma distribution

$$\sigma_z^2 \sim Inverse - Gamma(a_z, b_z)$$

where a_z and b_z are chosen to be very small. The full conditional distribution for the variance σ_z^2 borrows information form the observed values with the assumption, the prior distribution, that is relatively small. As a result the posterior distribution is:

$$f(\sigma_z^2|Z, \rho, \mu, Z_0) \propto \prod_{t=1}^{N_t} f(Z_t|\rho, \mu, \sigma_z^2) \times f(Z_0|\rho, \mu, \sigma_z^2) \times f(\sigma_z^2)$$

Specifically, it follows

$$f(\sigma_z^2|.) \propto (\sigma_\theta^2)^{-a_z - \frac{N_t+1}{2}} exp\left\{ - \frac{\sum_{t=1}^{N_t} (Z_t - \rho Z_{t-1})^2 + (Z_0 - \mu)^2(1-\rho^2) + 2b_z}{2\sigma_z^2} \right\}$$

So given all other variables, σ_z^2 follows the distribution as follows:

$$\sigma_z^2|. \sim Inv - Gam\left(a_z + \frac{N_t + 1}{2}, b_z + \frac{\sum_{t=1}^{N_t} (Z_t - \rho Z_{t-1})^2 + (Z_0 - \mu)^2(1-\rho^2)}{2} \right)$$

6.4 Spatio-temporal model

We now give details of the MCMC for the spatio-temporal model presented in Section 7.3. The spatio-temporal exposure model is presented below.

$$
\begin{aligned}
ln(Y_{st}) &\sim N(Z_t + m_s, \sigma_\epsilon^2) \quad for \quad t = 1, \ldots, N_t, \\
Z_t &\sim N(\mu + \rho(Z_{t-1} - \mu), \sigma_z^2) \\
m_s &\sim MVN(0, \sigma_m^2 \Sigma(\phi)) \\
\sigma_\epsilon^2 &\sim Inverse - Gamma(a_\epsilon, b_\epsilon) \\
\sigma_z^2 &\sim Inverse - Gamma(a_z, b_z) \\
\sigma_m^2 &\sim Inverse - Gamma(a_m, b_m) \\
\phi &\sim Discrete - Uniform(a_\phi, b_\phi) \\
\rho &\sim Uniform(a_\rho, b_\rho) \\
\mu &\sim N(0, \psi^2)
\end{aligned}
$$

ρ, μ, σ_m^2, σ_z^2 and σ_ϵ^2 are assigned priors as in the measurement error model described in Section 6.3. The prior of ϕ is chosen as a discrete uniform distribution with upper and lower limits which constrain the spatial correlation to be within a reasonable range. The probability density function of Y_{st} (on the log scale) given all the other parameters is given as:

$$
f(ln(Y_{st})) \propto \left(\sigma_\epsilon^2\right)^{-\frac{1}{2}} exp\left\{ - \frac{[ln(Y_{st}) - (Z_t + m_s)]^2}{2\sigma_\epsilon^2} \right\}
$$

The likelihood function for the underlying temporal process Z_t given the parameters involved is given as:

$$
f(Z_t) \propto \left(\sigma_z^2\right)^{-\frac{1}{2}} exp\left\{ - \frac{[Z_t - \mu - \rho(Z_{t-1} - \mu)]^2}{2\sigma_z^2} \right\}
$$

Furthermore, the likelihood function of the spatial effect m is a zero mean multivariate normal distribution as follows:

$$
f(m_s|.) = \frac{1}{\sqrt{(2\pi)^{N_s} |\sigma_m^2 \Sigma(\phi)|}} exp\Big\{ - \frac{1}{2} m_s^T \left(\sigma_m^2 \Sigma(\phi)\right)^{-1} m_s \Big\}
$$

Hence the joint posterior of $(m_s, \gamma, \rho, \theta, \sigma_\epsilon^2, \sigma_z^2, \sigma_m^2, \phi)$ is given by:

$$
\begin{aligned}
f(m_s, \alpha, \rho, Z, \sigma_\epsilon^2, \sigma_z^2, \phi | ln(Y)) \quad \propto \quad & \prod_{s=1}^{N_s} \prod_{t=1}^{N_t} f(ln(Y_{st}) | m_s, \alpha, Z_t, \sigma_\epsilon^2) \\
& \times \prod_{t=1}^{N_t} f(Z_t | \rho, \sigma_z^2) \times \prod_{s=1}^{N_s} f(m_s | \phi, \sigma_m^2) \\
& \times f(Z_0 | \mu, \rho, \sigma_z^2) \times f(\alpha) \times f(\rho) \\
& \times f(\phi) \times f(\sigma_\epsilon^2) \times f(\sigma_z^2) \times f(\sigma_m^2)
\end{aligned}
$$

6.4.1 Sampling from the full conditional distribution of the correlation–distance parameter, ϕ

An important aspect when applying MCMC is the sampling methodology of the correlation-distance parameter ϕ. It can be highly problematic, since the only quantity ϕ relates to is spatial effect m which is random, meaning there is no information support from the data, therefore the posterior distribution of ϕ is largely determined by the prior, this leads to that ϕ often does not converge and the acceptance ratio is very low. In addition, if the prior distribution of ϕ is a Uniform distribution, the full conditional distribution can not be found in closed form and a Metropolis-Hastings algorithm is applied.

Berger et al. (2001) and Zhang (2004) have shown that the specification of a flat or continuous uniform prior with a large range leads to either an improper or nearly uniform posterior over the range of the prior, neither of which is really desirable. The discrete method is straightforward and offers a much faster algorithm because many of the required calculations can be done beforehand (Diggle et al., 1998; Sahu et al., 2007). Despite the possible loss of information when approximating a continuous distribution by a discrete one, it is also likely to give better mixing to the posterior distribution. Therefore, we adopt a Discrete-Uniform distribution as the prior for correlation–distance parameter ϕ in this thesis. It is given as

$$\phi \sim Discrete - Uniform(a_\phi, b_\phi)$$

The discrete uniform distribution is a probability distribution in which a finite number of values are equally likely to be observed. In order to construct the discrete uniform distribution here, we choose the lower limit a_ϕ and b_ϕ for a sequence of a finite number of values which have equal space between them, hence a finite number of outcomes are equally likely to occur. Since ϕ is only related to the spatial effects m, the posterior distribution of it can be written as

$$f(\phi | m, \sigma_m^2) \propto f(m | \phi, \sigma_m^2) \times f(\phi)$$

The density of posterior of ϕ is not a closed form of any standard distribution because of the likelihood of m and therefore the Metropolis-Hastings algorithm is applied. The acceptance ratio function is only a function of the likelihood function of m, in this context, and so the proposed value of ϕ is accepted with probability

$$c = min\left\{1, \frac{f(m|\phi^{\star}, \sigma_m^2)}{f(m|\phi^{c}, \sigma_m^2)}\right\}$$

where $\phi^{\star}$ is the proposed value and ϕ^{c} denotes the current value in a certain iteration of MCMC. If the denominator, which is the likelihood of m based on the proposed ϕ is very small, then this ratio function produces extreme large values, hence it may cause numerical problems, therefore the calculation is performed on the log scale. The log acceptance rate is as follows

$$c^{\star} = min\left\{0, log(f(m|\phi^{\star}, \sigma_m^2)) - log(f(m|\phi^{c}, \sigma_m^2))\right\}$$

6.4.2 Sampling from the full conditional distribution of the spatial effects, m

The likelihood function of the spatial effects m is a zero mean multivariate normal distribution.

$$f(m|\sigma_m^2, \phi) \sim MVN(0, \sigma_m^2 \Sigma_{\phi})$$

hence in specific form:

$$f(m|.) = \frac{1}{\sqrt{(2\pi)^{N_s}|\sigma_m^2 \Sigma_{\phi}|}} exp\Big\{ -\frac{1}{2} m^T \big(\sigma_m^2 \Sigma_{\phi}\big)^{-1} m \Big\}$$

One method of sampling m_s is by generating all values simultaneously from the multivariate normal distribution. However this may lead to poor acceptance rates of ϕ. This is because the simultaneous sampling may produce a total difference between two vectors of m which is considerably large, consequently, the chance of accepting this proposed ϕ is reduced at each iteration according to the proposal ratio function in Metropolis-Hastings algorithm.

Here we perform separate updating for each element of m conditional on all other elements. In order to simplify the notation, we denote $\sigma_m^2 \Sigma_{\phi}$ by $\mathbf{\Sigma}$, and use the conditional properties of the multivariate normal distribution. We obtain the conditional likelihood function of each m_s as follows:

$$
\begin{aligned}
m_s &\sim N\big(\overline{\mu}, \overline{\Sigma}\big) \\
\overline{\mu} &= \mathbf{\Sigma}_{\{s,-s\}}\mathbf{\Sigma}^{-1}_{\{-s,-s\}}m_{-s} \\
\overline{\Sigma} &= \mathbf{\Sigma}_{\{s,s\}} - \mathbf{\Sigma}_{\{s,-s\}}\mathbf{\Sigma}^{-1}_{\{-s,-s\}}\mathbf{\Sigma}_{\{-s,s\}}
\end{aligned}
$$

where $\mathbf{\Sigma}_{\{s,-s\}}$ represents the vector which is taken from the s^{th} row of matrix $\mathbf{\Sigma}$ without the s^{th} element in the vector, while $\mathbf{\Sigma}_{\{-s,-s\}}$ represents the matrix $\mathbf{\Sigma}$ without s^{th} row and column, in addition $\mathbf{\Sigma}_{\{s,s\}}$ is an entry of matrix $\mathbf{\Sigma}$ where the s^{th} row and the s^{th} column interact. Hence, the full conditional distribution of a single m_s with respect to location s borrows information from the corresponding set of observed air pollution measurements at the same location. It follows that

$$
f(m_s|Y, Z, \phi, \sigma^2_\epsilon, \sigma^2_m) \propto \prod_{s=1}^{N_s} f(ln(Y_{st})|Z_t, m_s, \sigma^2_\epsilon) \times f(m_s|\phi, \sigma^2_m)
$$

and in this case the likelihood function of $f(ln(Y_{st}) \mid \theta, m_s, \sigma^2_\epsilon)$ with respect to m_s can be written as:

$$
f(ln(Y_{st})|Z_t, m_s, \sigma^2_\epsilon) \propto exp\Bigg\{ - \frac{\sum_{t=1}^{N_t}\big[m_s^2 - 2(ln(Y_{st}) - Z_t)\big]}{2\sigma^2_\epsilon} \Bigg\}
$$

Therefore, the full conditional distribution for each m_s is:

$$
f(m_s|.) \propto exp\Bigg\{ - \frac{m_s^2 - 2\big(\frac{\sum_{t=1}^{N_t}(ln(Y_{st}) - Z_t)\overline{\Sigma} + \overline{\mu}\sigma^2_\epsilon}{\sigma^2_\epsilon + N_t\overline{\Sigma}}\big)m_s}{2\Big(\frac{\overline{\Sigma}\sigma^2_\epsilon}{\sigma^2_\epsilon + N_t\overline{\Sigma}}\Big)} \Bigg\}
$$

This is the kernel of a normal distribution, hence the posterior distribution of m_s is:

$$
m_s \sim N\Bigg(\frac{\sum_{t=1}^{N_t}(ln(Y_{st}) - Z_t)\overline{\Sigma} + \overline{\mu}\sigma^2_\epsilon}{\sigma^2_\epsilon + N_t\overline{\Sigma}}, \frac{\overline{\Sigma}\sigma^2_\epsilon}{\sigma^2_\epsilon + N_t\overline{\Sigma}}\Bigg)
$$

6.4.3 Sampling from the full conditional distribution of the underlying trend, Z

The values of the underlying level, Z_t, have distribution function

$$
f(Z|\mu, \rho, \sigma^2_z) = \prod_{t=1}^{N_t} f(Z_t|Z_{t-1}, \mu, \rho, \sigma^2_z)
$$

Each of the terms in the product can be written as

$$f(Z_t|Z_{t-1},\mu,\rho,\sigma_z^2) = \frac{1}{\sqrt{2\pi\sigma_z^2}} exp\Bigg\{ - \frac{[Z_t - \mu - \rho(Z_{t-1}-\mu)]^2}{2\sigma_z^2} \Bigg\}$$

Information about Z_0 is required in order to obtain the full conditional distribution. Therefore, a prior is assigned to it:

$$f(Z_0|\mu,\rho,\sigma_z^2) = \frac{\sqrt{(1-\rho^2)}}{\sqrt{2\pi\sigma_z^2}} exp\Bigg\{ - \frac{(Z_0 - \mu)^2(1-\rho^2)}{2\sigma_z^2} \Bigg\}$$

Hence the full conditional distribution of Z_t for $t = 1, \ldots, N_t - 1$ can be written as

$$\begin{aligned} f(Z_t|Y_t, m, \mu, \rho, Z_{t-1}, Z_{t+1}, \sigma_z^2) \quad \propto \quad & \prod_{s=1}^{N_s} f(ln(Y_{st})|Z_t, m_s) \times f(Z_t|\mu,\rho,Z_{t-1},\sigma_z^2) \\ & \times f(Z_{t+1}|\mu,\rho,Z_t,\sigma_z^2) \end{aligned}$$

which is

$$f(Z_t|.) \propto exp\Bigg\{ - \frac{Z_t^2 - 2\Big\{ \frac{\sigma_z^2 \sum_{s=1}^{N_s}(ln(Y_{st}) - m_s) + \sigma_\epsilon^2[\rho(Z_{t-1}+Z_{t+1}) + \mu(1-\rho)^2]}{N_s\sigma_z^2 + (1+\rho^2)\sigma_\epsilon^2 + \sigma_z^2\sigma_\epsilon^2} \Big\} Z_t}{\frac{2\times\sigma_z^2\sigma_\epsilon^2}{N_s\sigma_z^2 + (1+\rho^2)\sigma_\epsilon^2 + \sigma_z^2\sigma_\epsilon^2}} \Bigg\}$$

This is the closed form of a normal distribution, hence each Z_t can be updated using a Gibbs sampler from the posterior distribution

$$\begin{aligned} Z_t|. \sim \quad N\Bigg(& \frac{\sigma_z^2 \sum_{s=1}^{N_s}(ln(Y_{st}) - m_s) + \sigma_\epsilon^2[\rho(Z_{t-1}+Z_{t+1}) + \mu(1-\rho)^2]}{N_s\sigma_z^2 + (1+\rho^2)\sigma_\epsilon^2 + \sigma_z^2\sigma_\epsilon^2}, \\ & \frac{\sigma_z^2\sigma_\epsilon^2}{N_s\sigma_z^2 + (1+\rho^2)\sigma_\epsilon^2 + \sigma_z^2\sigma_\epsilon^2} \Bigg) \end{aligned}$$

The special cases are the posterior distributions of Z_0 and Z_{N_t}, the former of which borrows information from Z_1, while the latter involves only the value before it which is Z_{N_t-1}, hence they can be written as:

$$
\begin{aligned}
Z_{N_t}|. \sim \quad & N\Bigg(\quad \frac{\sigma_z^2 \sum_{s=1}^{N_s}(ln(Y_{sN_t}^{(2)}) - m_s) + \sigma_\epsilon^2[\rho Z_{t-1} + \mu(1-\rho)]}{N_s\sigma_z^2 + \sigma_\epsilon^2 + \sigma_z^2\sigma_\epsilon^2}, \\
& \qquad \frac{\sigma_z^2\sigma_\epsilon^2}{N_s\sigma_z^2 + \sigma_\epsilon^2 + \sigma_z^2\sigma_\epsilon^2}\Bigg) \\
Z_0|. \quad \sim \quad & N\Big(\mu + \rho(Z_1 - \mu), \sigma_z^2\Big)
\end{aligned}
$$

6.4.4 Sampling from the full conditional distribution of the autoregressive process intercept, μ

The prior for the intercept term μ is assumed to be vague and to be represented by a Gaussian distribution with large variance

$$\mu \sim N(0, \psi^2) \propto exp\Big\{ - \frac{\mu^2}{2 \times \psi^2}\Big\}$$

The full conditional distribution of μ only depends on the autoregressive process and the priors, it can be written as:

$$f(\mu|Z, \rho, \sigma_z^2) \sim \prod_{t=1}^{N_t} f(Z_t|Z_{t-1}, \mu, \rho, \sigma_z^2) \times f(Z_0|\mu, \rho, \sigma_z^2) \times f(\mu)$$

Therefore the full conditional distribution for μ is:

$$\mu|. \sim N\left(\frac{\psi^2[(1-\rho)\sum_{t=1}^{N_t}(Z_t - \rho Z_{t-1}) + (1-\rho^2)Z_0]}{\sigma_z^2 + \psi^2(1-\rho^2) + \psi^2 N_t(1-\rho)^2}, \frac{\psi^2\sigma_z^2}{\sigma_z^2 + \psi^2(1-\rho^2) + \psi^2 N_t(1-\rho)^2}\right)$$

6.4.5 Sampling from the full conditional distribution of the autoregressive process parameter, ρ

The autoregressive process parameter ρ is assigned a continuous uniform prior which is given as

$$\rho \sim Uniform(a_\rho, b_\rho) \propto 1$$

Since the information on Z_0 is involved in the full conditional distribution of ρ, we can not obtain the posterior distribution directly. In this case, the Metropolis Hastings algorithm is adopted here to generate from the full conditional distribution of ρ. In order to simplify the process, the proposal distribution is chosen as:

$$\rho|Z,\mu,\sigma_z^2,Z_0 \sim N\left(\frac{\sum_{t=1}^{N_t}(Z_{t-1}-\mu)(Z_t-\mu)}{\sum_{t=1}^{N_t}(Z_{t-1}-\mu)^2}, \frac{\sigma_z^2}{\sum_{t=1}^{N_t}(Z_{t-1}-\mu)^2}\right)$$

In this circumstance, the acceptance ratio only depends on the density function of Z_0, therefore, the proposed value of ρ is accepted with probability:

$$c = min\left\{1, \frac{f(Z_0|\rho^\star,\mu,\sigma_z^2)}{f(Z_0|\rho^c,\mu,\sigma_z^2)}\right\}$$

where $\rho^\star$ denotes the proposed value of ρ, and ρ^c represents the current value of ρ.

6.4.6 Sampling from the full conditional distribution of the variance of measurement error, σ_ϵ^2

The hyper-prior for the variance of measurement error σ_ϵ^2 is a Gamma distribution

$$\sigma_\epsilon^2 \sim Inverse - Gamma(a_\epsilon, b_\epsilon)$$

The posterior distribution is:

$$f(\sigma_\epsilon^2|ln(Y),Z,m) \propto \prod_{t=1}^{N_t}\prod_{s=1}^{N_s} f(ln(Y_{st})|Z_t,m_s,\sigma_\epsilon^2) \times f(\sigma_\epsilon^2)$$

which is of the form

$$f(\sigma_\epsilon^2|.) \propto (\sigma_\epsilon^2)^{-a_\epsilon-\frac{N_sN_t}{2}} exp\left\{-\frac{\sum_{t=1}^{N_t}\left[\sum_{s=1}^{N_s}(ln(Y_{st})-(Z_t+m_s))^2\right]+2b_\epsilon}{2\sigma_\epsilon^2}\right\}$$

This is a kernel of Inverse-Gamma distribution which is represented as:

$$\sigma_\epsilon^2|. \sim Inv - Gam\left(a_\epsilon + \frac{N_sN_t}{2}, b_\epsilon + \frac{\sum_{t=1}^{N_t}\left[\sum_{s=1}^{N_s}(ln(Y_{st})-(Z_t+m_s))^2\right]}{2}\right)$$

6.4.7 Sampling from the full conditional distribution of variance of autoregressive process, σ_z^2

The hyper-prior for the variance of measurement error σ_z^2 is a Gamma distribution

$$\sigma_z^2 \sim Inverse - Gamma(a_z, b_z)$$

The posterior distribution is:

$$f(\sigma_z^2|Z,\rho,\mu,Z_0) \propto \prod_{t=1}^{N_t} f(Z_t|\rho,\mu,\sigma_z^2) \times f(Z_0|\rho,\mu,\sigma_z^2) \times f(\sigma_z^2)$$

Specifically, it follows

$$f(\sigma_z^2|.) \propto (\sigma_\theta^2)^{-a_\theta-\frac{N_t+1}{2}} exp\Bigg\{ - \frac{\sum_{t=1}^{N_t}(Z_t-\rho Z_{t-1})^2 + (Z_0-\mu)^2(1-\rho^2) + 2b_z}{2\sigma_z^2} \Bigg\}$$

So σ_z^2 given all other variables follows the distribution:

$$\sigma_z^2|. \sim Inv-Gam\left(a_z + \frac{N_t+1}{2}, b_z + \frac{\sum_{t=1}^{N_t}(Z_t-\rho Z_{t-1})^2 + (Z_0-\mu)^2(1-\rho^2)}{2}\right)$$

6.4.8 Sampling from the full conditional distribution of the spatial variance, σ_m^2

The hyper-prior for the variance of the measurement error σ_m^2 is a Gamma distribution

$$\sigma_m^2 \sim Inverse-Gamma(a_m, b_m)$$

The posterior distribution is:

$$f(\sigma_m^2|m,\phi) \propto f(m|\phi,\sigma_m^2) \times f(\sigma_m^2)$$

Hence it follows the density function

$$f(\sigma_m^2|.) \propto (\sigma_m^2)^{-a_m-\frac{N_s}{2}} exp\Big\{ - \frac{1}{2} m_s^T \big(\Sigma(\phi)\big)^{-1} m_s - b_m \Big\}$$

This is a kernel of Inverse-Gamma distribution, therefore, σ_m^2 given all the other parameters follows the distribution:

$$\sigma_m^2|. \sim Inv-Gam\left(a_m + \frac{N_s}{2}, b_m + \frac{1}{2} m_s^T \big(\Sigma(\phi)\big)^{-1} m_s\right)$$

Chapter 7

Assessment of spatio-temporal and measurement error exposure models for estimating health risks

In this chapter, we examine the effectiveness of the spatial-temporal model and the measurement error model in dealing with the simulated data comprising both spatial variation and measurement error, by investigating the accuracy of estimated relative risk compared to the true value. The models are implemented by MCMC in this simulation study. Considering the heavy computational burden of running each MCMC and the fact that we may need to run the models a large number of times, parallel computing is used. It is possible to perform parallel computing using R which allows us to take advantage of modern multi-core hardware (McCallum and Weston, 2011). The following are the details of this simulation study. First of all, the method used to generate the data is presented. This is followed by a description of two models for modelling exposures: a spatial–temporal model and a measurement error model and the results from the simulation study using them both.

7.1 Data generation

The data generation procedure follows the one described in Section 5.5.1. For each dataset, in order to generate both spatial variation and measurement error in the data, the model used to generate the simulated data is as follows,

$$
\begin{aligned}
Y_t^{(1)} &\sim Poisson(\mu_t) \quad for \;\; t = 1, \ldots, N_t, \\
ln(\mu_t) &= \beta_1 exp(Z_t^{(2)}) \\
ln(Y_{st}^{(2)}) &\sim N(Z_t^{(2)} + m_s, \sigma_\epsilon^2) \\
Z_t^{(2)} &\sim N(\mu^{(2)} + \rho(Z_{i-1}^{(2)} - \mu^{(2)}), \sigma_z^2) \\
m_s &\sim MVN(0, \sigma_m^2 \Sigma_\phi)
\end{aligned}
\tag{7.1}
$$

where $Y_t^{(1)}$ represents mortality, β_1 is the log relative risk. In the exposure model, $Y_{st}^{(2)}$ denotes the measured pollutant levels, $Z^{(2)}$ represents true underlying exposures, and $\mu^{(2)}$ is the intercept term in the temporal process.

For each dataset, air pollution are generated for 365 consecutive days (1 year) at 20 locations in the study area; these locations are shown in Figure (5-5). Since we assume the pollution data has a stationary model without long-term trend, an AR(1) process with mean level $\mu^{(2)}$ is applied to generate underlying true exposure $Z^{(2)}$ which is specified by $\rho = 0.8$ and $\sigma_z = 0.25$. In addition, since our focus is the relative risk of increase of 10 units of pollution, the intercept term $\mu^{(2)}$ is chosen as 1.38 ($ln(40/10)$).

Then the mortality data $Y^{(1)}$ is generated using only the underlying exposures and the corresponding log relative risk. The individual level association between ambient pollution levels and mortality is fixed at a relative risk of 20 percent for an increase in ten units of pollution, therefore, here $\beta_1 = log(1.2)$. The mortality data follow poisson distributions with log mean level equal to the product of $Z^{(2)}$ and β_1. In particular, for clarity, all mortalities are assumed to be only caused by true air pollution exposure and hence there is no intercept term.

The set of spatial random effects, $\mathbf{m}$, is generated from multivariate normal distribution with an exponential covariance function, in which the correlation–distance parameter, ϕ, is chosen to be either 0.121 or 0.015, representing low and high spatial correlation respectively. In addition, eight different values of the spatial standard deviation σ_m are used; from 0.1 to 1. Measurement error is assumed to arise from a normal distribution with zero mean and σ_ϵ assigned one of eight values from 0.1 to 1. The measured air pollution data $Y^{(2)}$ are then simulated by adding spatial variation m and measurement error to the true underlying exposure level $Z^{(2)}$.

7.2 Parallel computing

The traditional software is written for serial computation, such that a problem is broken into a discrete series of instructions that are executed on a single processor sequentially one after another. In fact many problems are so large or complex that it is impractical to solve them serially, especially given limited computational power. Parallel computing use multiple compute resources simultaneously to solve a computational problem which is broken into discrete parts that can be executed concurrently on different processors, hence the problem is solved in less time with multiple compute resources than with a single compute resource. Compared to serial computing, parallel computing is much better suited for modeling and simulations.

The compute resources used for parallel computing are typically a single computer with multiple cores or an arbitrary number of such computers connected by a network. Modern computers, even laptops, are parallel in architecture with multiple processors, and parallel software is specifically intended for the hardware with multiple cores. In fact the serial programs run on modern computers waste potential computing power.

Here we use the package in R called 'parallel' and work on a 4-core computer. In this simulation study, since each simulation scenario generates 200 datasets, MCMC is implemented on 8 datasets simultaneously each time (multiple times of number of cores), and it is run 25 times. Also, we run multiple Markov chains instead one and combine the samples at the end. More details of parallel computing in R can be found in McCallum and Weston (2011).

7.3 Assessment of the spatio-temporal model

To assess the effectiveness of the spatio-temporal model we examine the estimated relative risks and their coverage probabilities using the spatio-temporal model for modelling exposures. As described in Chapter 5, we use a two–stage approach using multiple datasets generated from the (joint) posterior distributions from this model, from the MCMC samples, in a health model, with the results being combined using multiple imputation. In this case 200 such datasets were generated and two health models used with Poisson and quasi-likelihoods. The spatio-temporal exposure model used in the simulation studies is as follows;

$$
\begin{aligned}
Y_t^{(1)} &\sim Poisson(\mu_t) \quad for \;\; t = 1, \ldots, N_t, \\
ln(\mu_t) &= \beta_1 \exp(Z_t^{(2)}) \\
ln(Y_{st}^{(2)}) &\sim N(Z_t^{(2)} + m_s, \sigma_\epsilon^2) \\
Z_t^{(2)} &\sim N(\mu^{(2)} + \rho(Z_{t-1}^{(2)} - \mu^{(2)}), \sigma_z^2) \\
m_s &\sim MVN(0, \sigma_m^2 \Sigma(\phi)) \\
\sigma_\epsilon^2 &\sim Inverse - Gamma(0.01, 0.01) \\
\sigma_z^2 &\sim Inverse - Gamma(0.01, 0.01) \\
\sigma_m^2 &\sim Inverse - Gamma(0.01, 0.01) \\
\phi &\sim Discrete - Uniform(0.0025, 0.1151) \\
\rho &\sim Uniform(0, 1) \\
\mu^{(2)} &\sim N(0, 10^3)
\end{aligned}
\tag{7.2}
$$

7.3.1 Results

Tables 7.1 and Table 7.2 show the results from the 200 health analyses, each using datasets comprising predictions from the spatio–temporal exposure model. Table 7.1 gives the results for Poisson health models and Table 7.2 for the quasi-likelihood versions. Overall estimates of relative risks are shown together with associated 95% confidence intervals. Coverage probabilities, showing the proportion of times the 95% CIs from the individual analyses contain the true value are also given. In each table, 8 results from the simulation scenarios are presented, in which the smallest value for both measurement error standard deviation and spatial standard deviation are chosen as 0.05. In order to examine the efficiency of the model in extreme circumstances, we use spatial standard deviation and measurement error standard deviation as high as 1.

From Table 7.1, we can see that the estimates of relative risks are all very close to the true value 1.2. When the measurement error standard deviation and spatial standard deviation are both very small, e.g. 0.05, the 95% CIs are narrow and the coverage probabilities are as high as 90 percent. When the measurement error standard deviation and spatial standard deviation increase the confidence intervals become wider and the coverage probabilities go down. Eventually, when both standard deviations reach 1, the intervals are widest and the coverage probabilities are only 17 percent and 8 percent when the distance–correlations parameter is $\phi = 0.121$ and $\phi = 0.015$ respectively.

When spatial correlation is high ($\phi = 0.015$), then the intervals are generally wider and the coverage probabilities are lower than those estimated when spatial correlation is low. For example, when the measurement error standard deviation and spatial standard deviation are both 0.1, the 95% CI from the high spatial correlation scenario is (1.157, 1.244) which is wider than the one seen when there is low spatial correlation, (1.183, 1.220). In addition, the coverage probability in the former case is 37 percent which is much lower than 88 percent seen in the low spatial correlated datasets.

The results in Table 7.2 tell a similar story to those in Table 7.1. Comparing the results between the two tables, we can see the 95% CI intervals in Table 7.2 are wider than the corresponding ones in Table 7.1 for the same level of measurement error and spatial standard deviation. Moreover, the coverage probabilities are generally higher than in the Poisson case. For example, when the measurement error standard deviation and spatial standard deviation are equal to 1, the 95% CI in Table 7.2 for the high spatial correlation scenario is (1.039, 2.509) which is wider than (1.037, 2.371) in Table 7.1. Furthermore, the coverage probability is 11 percent in this case while the one shown in Table 7.1 is 8 percent. However it is noted that these levels of spatial variability are extremely high and would certainly be well beyond what might be observed in real life.

Sensitivity analysis was conducted to ascertain whether the choice of priors affected the results. These analysis focused on prior specification for the variance and correlation parameters, because the other parameters are given standard priors that should work well in a variety of situations. A series of Inverse-Gamma(ϵ_1, ϵ_2) priors (where $\epsilon_i = 0.1, 0.01, 0.001$) were specified for each variance parameter. It appeared that these choices had little effect on the resulting posteriors.

7.4 Assessment of the measurement error model

Compared to the spatio-temporal model, the measurement error model provides a somewhat simpler and less computationally demanding alternative to the spatio–temporal model for exposures. If there are both spatial variation and measurement error in the air pollution data, the measurement error model will simply treat them both as random error without structure. The measurement error model used in the simulation study is as follows;

Spatial decay parameter	ME sd	SP sd	Results
ϕ=0.121	0.05	0.05	1.194 (1.184, 1.204) 91%
	0.05	0.1	1.198 (1.178, 1.219) 89%
	0.1	0.1	1.199 (1.183, 1.220) 88%
	1	1	1.195 (1.082, 1.550) 17%
ϕ=0.015	0.05	0.05	1.200 (1.181, 1.217) 84%
	0.05	0.1	1.210 (1.173, 1.249) 48%
	0.1	0.1	1.197 (1.157, 1.244) 37%
	1	1	1.193 (1.037, 2.371) 8%

Table 7.1: Overall measures of risk and 95% confidence intervals, together with coverage probabilities, based on 200 datasets based on the spatio–temporal exposure model. Results are for Poisson health models under different exposure scenarios where 'ME sd' stands for 'measurement error standard deviation', and 'SP sd' means 'spatial standard deviation' in the simulated data.

Spatial decay parameter	ME sd	SP sd	Results
ϕ=0.121	0.05	0.05	1.201 (1.197, 1.216) 93%
	0.05	0.1	1.202 (1.186, 1.221) 90%
	0.1	0.1	1.206 (1.196, 1.238) 79%
	1	1	1.192 (1.060, 1.491) 25%
ϕ=0.015	0.05	0.05	1.194 (1.179, 1.217) 86%
	0.05	0.1	1.199 (1.171, 1.246) 57%
	0.1	0.1	1.201 (1.170, 1.343) 48%
	1	1	1.198 (1.039, 2.509) 11%

Table 7.2: Overall measures of risk and 95% confidence intervals, together with coverage probabilities, based on 200 datasets based on the spatio–temporal exposure model. Results are for quasi–likelihood health models under different exposure scenarios where 'ME sd' stands for 'measurement error standard deviation', and 'SP sd' means 'spatial standard deviation' in the simulated data.

$$
\begin{aligned}
Y_t^{(1)} &\sim Poisson(\mu_t) \quad for \;\; t = 1, \ldots, N_t, \\
ln(\mu_t) &= \beta_1 exp(Z_t^{(2)}) \\
ln(Y_{st}^{(2)}) &\sim N(Z_t^{(2)}, \sigma_\epsilon^2) \qquad (7.3) \\
Z_t^{(2)} &\sim N(\mu^{(2)} + \rho(Z_{t-1}^{(2)} - \mu^{(2)}), \sigma_z^2) \\
\sigma_\epsilon^2 &\sim Inverse - Gamma(0.01, 0.01) \\
\sigma_z^2 &\sim Inverse - Gamma(0.01, 0.01) \\
\rho &\sim Uniform(0, 1)
\end{aligned}
\qquad (7.4)
$$

As with the spatio–temporal model, 200 datasets were generated for each of the different combinations of spatial and measurement error standard deviations and spatial correlation. Using each dataset the MCMC implementation of measurement error exposure model is run for at least 20000 iterations. For each scenario, the 200 generated datasets of exposure were used in 200 health analyses and the individual results combined to obtain an overall measure or relative risk together with 95% confidence intervals. These confidence intervals are used to calculate coverage probabilities.

7.4.1 Results

Tables 7.3 and Table 7.4 give the overall estimates of relative risks and 95% confidence intervals using exposure estimates from the measurement error model using Poisson and quasi-likelihood health models respectively. As with the spatio–temporal models, the results for eight combinations of measurement error and spatial standard deviations and spatial correlation are given together with coverage probabilities.

As when the spatio–temporal model was used to model the underlying exposures, from Table 7.3 we can see that the estimates of relative risks are all close to the true value of 1.2. However, again there are differences in the width of the CIs. When measurement error standard deviation and spatial standard deviation are both very small, e.g. 0.05, the 95% CIs are narrow, and the corresponding coverage probabilities are close to 90 percent for the quasi-likelihood example. The CIs become wider and the coverage probabilities decrease when the measurement error standard deviation and spatial standard deviation are increased. Eventually, when both standard deviation reach 1, the upper limit of the CI is very high and the coverage probabilities reduce to 15 percent and 5 percent for distance–correlation parameter $\phi = 0.121$ and $\phi = 0.015$ respectively. Furthermore, we can see if the spatial correlation is high ($\phi = 0.015$),

then the CIs are generally wider and coverage probabilities lower than when there is low spatial correlation.

The results in Table 7.4 are similar to those seen in Table 7.3 but with wider CIs in Table 7.4 than seen in Table 7.3 for the same level of measurement error standard deviation and spatial standard deviation. Moreover, the coverage probabilities are generally higher than the ones in the first table. For example, when measurement error standard deviation and spatial standard deviation both equal 1, the 95% CI in Table 7.4 for the high spatial correlation scenario is (1.048, 5.004) which is wider than the (1.027, 3.005) seen in Table 7.3.

Again, sensitivity analyses to the choice of priors was performed, as with the spatio–temporal model, and again little difference was seen in the posteriors.

7.5 Conclusion

In this chapter, the simulation studies assess the efficacy of spatial–temporal model and measurement error model in estimating relative risks. We found that both of these models are able to estimate the true risks to health in a much more accurate way than the standard model. Although spatial–temporal model provides narrower confidence intervals and higher coverage probabilities than measurement error model when facing the same level of spatial effect and measurement error in the data, the difference is rather small, hence there are advantages to use the less computationally demanding measurement error model. Furthermore, applying quasi-Poisson distribution widens the confidence intervals for both of these models, this implies that it is necessary to allow extra-Poisson variability for the mortality data in health models.

Spatial decay parameter	ME sd	SP sd	Results
ϕ=0.121	0.05	0.05	1.206 (1.188, 1.215) 83%
	0.05	0.1	1.211 (1.189, 1.232) 81%
	0.1	0.1	1.209 (1.188, 1.232) 63%
	1	1	1.208 (1.078, 1.641) 15%
ϕ=0.015	0.05	0.05	1.207 (1.186, 1.227) 60%
	0.05	0.1	1.197 (1.172, 1.243) 59%
	0.1	0.1	1.207 (1.177, 1.252) 42%
	1	1	1.173 (1.027, 3.005) 5%

Table 7.3: Overall measures of risk and 95% confidence intervals, together with coverage probabilities, based on 200 datasets based on the measurement error exposure model. Results are for Poisson health models under different exposure scenarios where 'ME sd' stands for 'measurement error standard deviation', and 'SP sd' means 'spatial standard deviation' in the simulated data.

Spatial decay parameter	ME sd	SP sd	Results
ϕ=0.121	0.05	0.05	1.203 (1.188, 1.212) 87%
	0.05	0.1	1.207 (1.186, 1.235) 84%
	0.1	0.1	1.198 (1.168, 1.218) 70%
	1	1	1.214 (1.082, 1.716) 17%
ϕ=0.015	0.05	0.05	1.203 (1.188, 1.233) 71%
	0.05	0.1	1.199 (1.172, 1.249) 66%
	0.1	0.1	1.202 (1.166, 1.251) 58%
	1	1	1.201 (1.048, 5.004) 9%

Table 7.4: Overall measures of risk and 95% confidence intervals, together with coverage probabilities, based on 200 datasets based on the measurement error exposure model. Results are for quasi–likelihood health models under different exposure scenarios where 'ME sd' stands for 'measurement error standard deviation', and 'SP sd' means 'spatial standard deviation' in the simulated data.

Chapter 8

Case study of the short–term effects of particulate matter on health in London

In this chapter, we apply the measurement error and spatio-temporal models described in Chapter 6 to a case–study of the effect of particulate matter on health in Greater London. We use the two–stage approach described in Section 4.2 to incorporate predictions from both the measurement error and spatio–temporal models (independently) into the health analysis.

The reminder of this chapter is organised as follows. The first section outlines the data used in this study. This is followed by a description of the standard model that is commonly used to analyse data of the sort considered here and the results of applying it. The next two sections present the results from implementing measurement error and spatio-temporal models to the London data.

8.1 Description of the data

The data used in this case study comprise daily observations of health and pollution data from the Greater London area between January 1st 2003 and December 31st 2005. The health data consist of daily counts of respiratory mortality which are only available in aggregate form for the entire area. The air pollution data comprise daily measurements of PM_{10} (particles smaller in size than $10\mu gm^{-3}$) from 112 monitoring sites located in Greater London area (as shown in Figure 8-1). Data collected form 'road side' and 'industry' locations are likely to have much higher concentrations, are likely to be extremely variable and are unlikely to represent the levels of pollution experienced by the population at risk. These are excluded from the analysis.

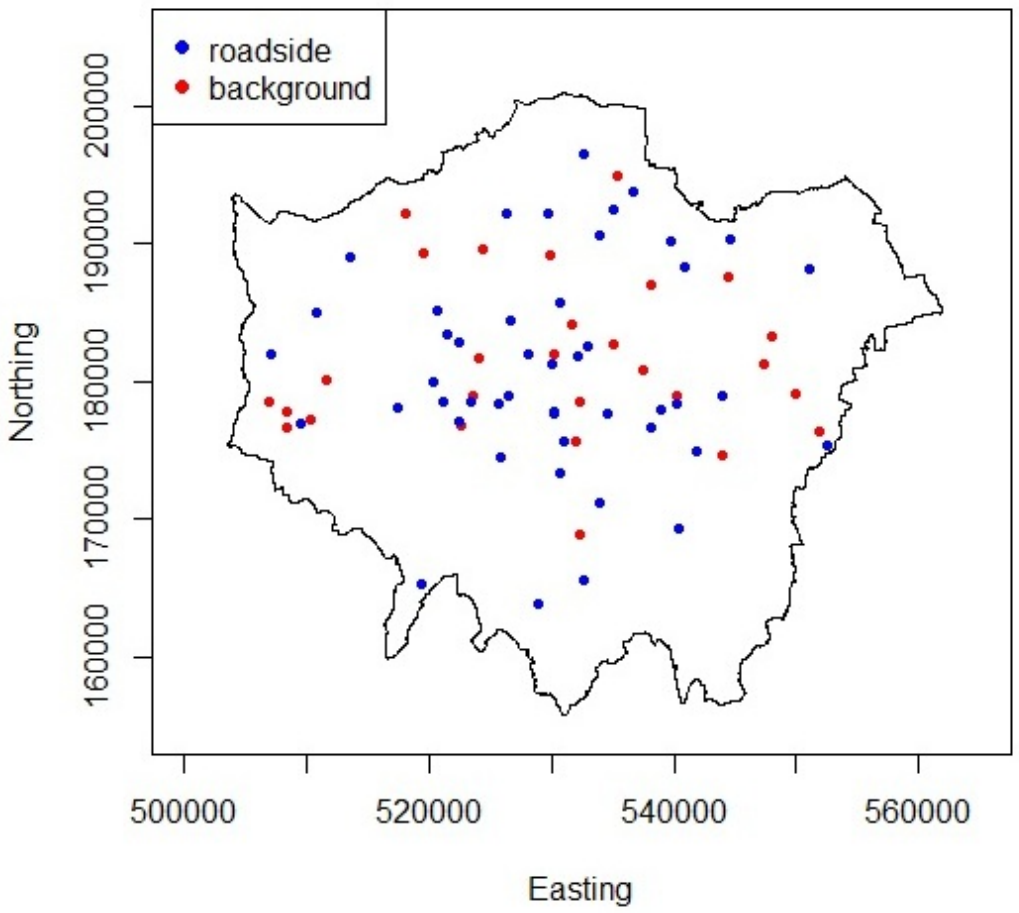

Figure 8-1: Locations of PM_{10} monitoring sites in Great London area from 2003 to 2005. Blue filled points denote roadside sites and the red points represent background sites.

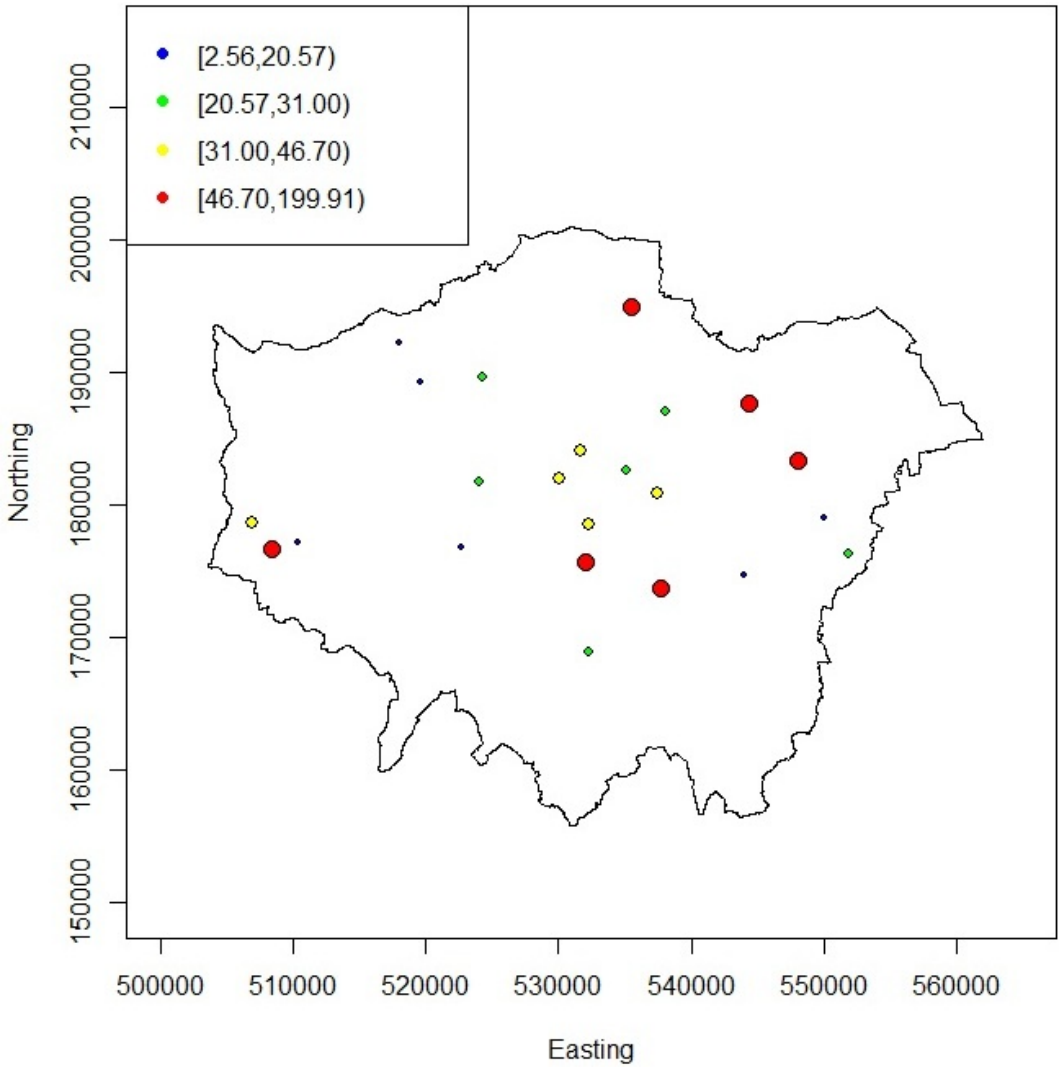

Figure 8-2: Locations of the pollution monitors used in the case–study, the solid circles represent the background sites with less than 25% missing measurements that we use in the study.

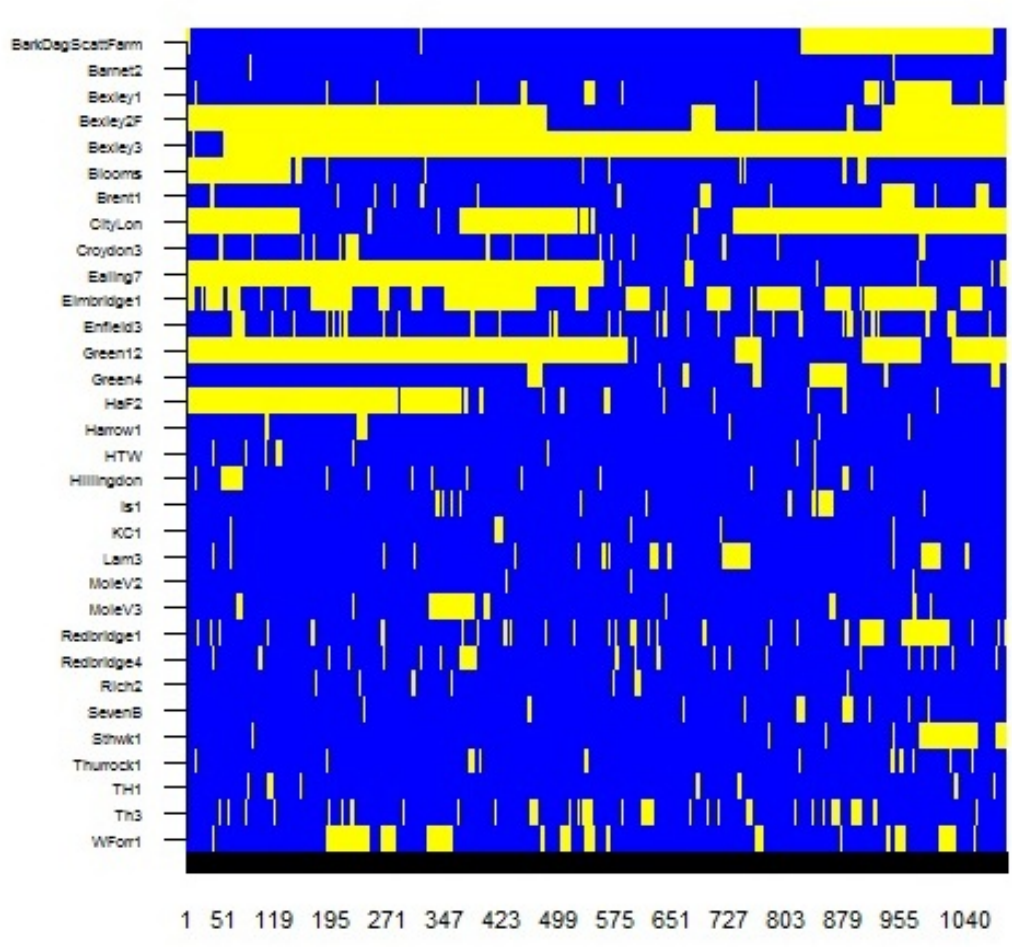

Figure 8-3: Schematic showing the days for which the monitoring sites returned daily data for the period 2003–2005. Each row represents a monitoring location with blue signifying that data was available and yellow showing periods of missing data.

Some of these sites include a large proportion of missing values (as shown in Figure 8-3), and sites with more than 25% missing observations over the study period are excluded. In addition, there are four sites located outside Great London area which will not be considered in the s-tudy. Applying these constraints leaves 23 monitoring locations which will be included within the case–study. Their locations can be seen in Figure 8-2.

Table 8.1: Estimates of the model parameters for the standard model together with 95% confidence intervals

Parameters	Estimated	95% CI
$exp(\beta_1)$	1.0298	1.0288 - 1.0307
β_0	2.9211	2.9190 - 2.9232

8.2 Standard model

In this section, the standard model is applied to the reduced subset of the London data. The standard exposure model involves simply averaging the air pollution measurements over all locations for each day. The health model is a Poisson GAM,

$$\begin{aligned} Y_t^{(1)} &\sim Poisson(\mu_t) \quad for \quad t = 1, \ldots, N_t, \\ ln(\mu_t) &= \beta_0 + \beta_1 \widetilde{Z}_{t-1}^{(2)} + S(time|\lambda_1) + S(temperature|\lambda_2) \\ \widetilde{Z}_t^{(2)} &= \overline{Y}_t^{(2)} = \frac{1}{N_s} \sum_{s=1}^{N_s} Y_{st}^{(2)} \end{aligned}$$

where $Y_t^{(1)}$ denotes the collected mortality on day t, and $\overline{Y}_t^{(2)}$ represents the averaged air pollution measurements over all locations on day t. The units are 10 μgm^{-3}. Here we use a lag of 1 in the exposures. The health model building process begins with removing the trend, seasonal variation and effects of temperature from the series of mortality data. These confounding factors are represented by penalised splines of calender time and temperature. In our model expression, the splines term is denoted by 'S', 'time' represents calender time in days, while 'temp' stands for the daily temperature. The intercept term is β_0 and the degrees of freedom for the splines for time and temperature are λ_1 and λ_2 respectively, each of which is estimated using MGCV (Wood, 2006). For the London data, the effective degrees of freedom for 'time' is 17 and is 7 for 'temperature'.

Table 8.1 shows estimates of the model parameters together with associated 95% confidence intervals. In particular, the main interest is the estimate of parameter β_1 which is the log of the relative risk, hence the estimated relative risk is $exp(0.0293) = 1.0298$. Therefore there is a significant increase in risk associated with an increase in PM_{10} on the previous day.

8.3 Spatio-temporal model

In this Section, we use the two-stage approach with a spatio-temporal model for the exposures (PM_{10}). A small amount of each pollution series in the data set is missing, which adds an additional complication to the analysis. In a traditional epidemiology analysis these might be ignored, however, using the underlying exposure model, predictions can be made for these locations and times. These predictions are made within the MCMC simulation and so samples from their posterior distributions can be used. The spatio-temporal model applied to the London data is given as

$$\begin{aligned} Y_t^{(1)} &\sim Poisson(\mu_t) \quad for \quad t = 1, \ldots, N_t, \\ ln(\mu_t) &= \beta_0 + \beta_1 Z_{t-1}^{(2)} + S(time|\lambda_1) + S(temperature|\lambda_2) \end{aligned}$$

$$\begin{aligned} ln(Y_{st}^{(2)}) &\sim N(Z_t^{(2)} + m_s, \sigma_\epsilon^2) \\ Z_t^{(2)} &\sim N(\mu^{(2)} + \rho(Z_{t-1}^{(2)} - \mu^{(2)}), \sigma_z^2) \\ m_s &\sim MVN(0, \sigma_m^2 \Sigma(\phi)) \\ \sigma_\epsilon^2 &\sim Inverse - Gamma(0.01, 0.01) \\ \sigma_z^2 &\sim Inverse - Gamma(0.01, 0.01) \\ \sigma_m^2 &\sim Inverse - Gamma(0.01, 0.01) \\ \phi &\sim Discrete - Uniform(0.0025, 0.1151) \\ \rho &\sim Uniform(0, 1) \\ \mu^{(2)} &\sim N(0, 10^3) \end{aligned} \tag{8.1}$$

The prior of parameter ρ is chosen as $Uniform(0, 1)$ as ρ represents the autoregressive process parameter. Inverse-Gamma $(0.01, 0.01)$ priors are assigned to σ_ϵ^2, σ_z^2 and σ_m^2. The correlation–distance parameter, ϕ, is given a discrete uniform prior with range (0.0025 , 0.1151), the details about prior choice of ϕ can be found in Section 6.4.1. The remaining parameters are assigned vague priors, using Gaussian distributions with large variances.

For the London data considered here, since the largest distance between two monitors is 56km, in order to restrict the spatial correlation to a reasonable range, the correlation–distance parameter ϕ is limited to be between 0.0025 and 0.1151. These values correspond to the spatial correlation at 54km being between 0.1 and 0.9. Details of the MCMC used to obtain samples from the posterior distributions of parameters and hyperparameters are given in Section 6.4.

Table 8.2: Estimates of the model parameters for the two-stage model with a spatio-temporal model for the exposures, together with their 95% confidence/credible intervals.

Parameters	Estimated	95% CI
$exp(\beta_1)$	1.0326	1.0131 - 1.0521
β_0	2.9175	2.8759 - 2.9592
ϕ	0.0904	0.0345 - 0.1462
ρ	0.6457	0.5998 - 0.6917
$\mu^{(2)}$	0.6572	0.6128 - 0.7016
σ_m^2	0.0686	0.0295 - 0.1668
σ_ϵ^2	0.0324	0.0318 - 0.0330
σ_z^2	0.0691	0.0631 - 0.0751

8.3.1 Results

Using the spatio-temporal exposure model, the MCMC algorithms were run for 40,000 iterations discarding the first 10,000 as 'burn in'. The estimates of the model parameters can be seen in Table 8.2. The estimated log scaled relative risk parameter β_1 is 0.0320, meaning the estimated relative risk is $exp(0.0320)$, that is 1.0326 (95% confidence interval 1.0131–1.0521). Figure 8-4 contains the trace plots of the parameters and shows that the samples of most of the posteriors seem to have converged and are relatively stable although there are some extreme values in the trace plot for the spatial variance and the correlation–distance parameter, ϕ, covers almost the entire range of the possible values. For the parameter ϕ it looks as though the samples are cut off at the maximum possible value.

Sensitivity analysis was conducted to ascertain whether the choice of priors affected the results. These analysis focused on prior specification for the variance and correlation parameters, because the other parameters are given standard priors that should work well in a variety of situations. A series of Inverse-Gamma (ϵ, ϵ) priors (where $\epsilon_i = 0.1, 0.01, 0.001$) were specified for each variance parameter. The results appeared to be robust to these changes. When other priors for ϕ were used, the feature seen in Panel (a) of Figure 8-4 was often repeated, e.g. the upper bound of the prior looked like an upper bound for the posterior. This is proved to be the case for increasing values of the upper bound which suggested that there is a possibility that the value of ϕ should be much greater, which would actually mean that there is a possibility that the spatial correlation is very low in these data. Furthermore, since ϕ and spatial variance σ_m^2 are closely related, the change of prior of ϕ also lead to a change in the estimate of σ_m^2. For example, when the upper bound of prior of ϕ is set to be 1, the estimated median of posterior distribution of ϕ is 0.5118, and the corresponding median of σ_m^2 is reduced to 0.0323. If the

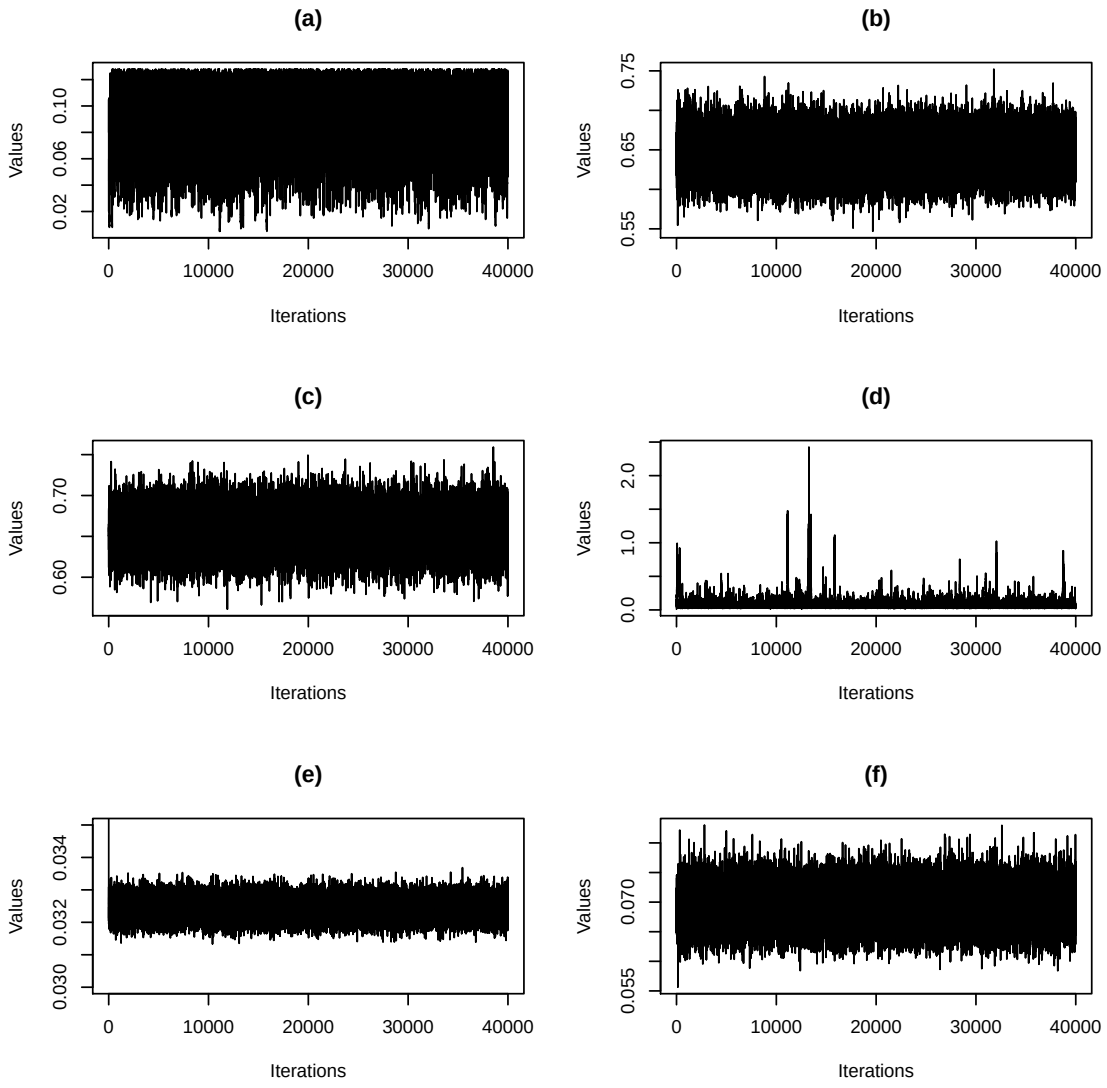

Figure 8-4: The trace plot of sampled posteriors of models parameters from the two-stage Bayesian spatio-temporal model implemented by MCMC with 40,000 iterations. (a):ϕ, (b):ρ, (c):$\mu^{(2)}$, (d):σ_m^2, (e):σ_ϵ^2, (f):σ_z^2.

upper bound of prior of ϕ is increased to 10, then the estimated medians of posteriors of ϕ and σ_m^2 are 4.8561 and 0.0304 respectively. For both these two cases, the estimations of all other parameters stay the same as the values shown in Table 8.2. This is because the deductive spatial effect m are extremely small such as 2.0319×10^{-6} and 6.0933×10^{-44} in these two cases if we consider the average distance between two monitoring locations in London to be $20km$. This also explains why in the second case, although not much reduction is shown in the estimated median of posterior of σ_m^2 compared to the dramatic increase of ϕ, it does not affect the results of other estimations as the first case.

8.4 Measurement error model

In this section we apply the measurement error model for the exposures in place of the spatio-temporal model used in Section 8.3. The measurement error model is as follows,

$$\begin{aligned} Y_t^{(1)} &\sim Poisson(\mu_t) \quad for \quad t = 1, \ldots, N_t, \\ ln(\mu_t) &= \beta_0 + \beta_1 Z_{t-1}^{(2)} + S(time|\lambda_1) + S(temperature|\lambda_2) \end{aligned}$$

$$\begin{aligned} ln(Y_{st}^{(2)}) &\sim N(Z_t^{(2)}, \sigma_\epsilon^2) \\ Z_t^{(2)} &\sim N(\mu^{(2)} + \rho(Z_{t-1}^{(2)} - \mu^{(2)}), \sigma_z^2) \\ \sigma_\epsilon^2 &\sim Inverse - Gamma(0.01, 0.01) \\ \sigma_z^2 &\sim Inverse - Gamma(0.01, 0.01) \\ \rho &\sim Uniform(0, 1) \\ \mu^{(2)} &\sim N(0, 10^3) \end{aligned} \tag{8.2}$$

Prior distributions for the parameters of the model are chosen as follows. The prior choice of parameter ρ is chosen as $Uniform(0, 1)$, this is because ρ represents the autoregressive process parameter. Inverse-Gamma $(0.01, 0.01)$ priors were used for both σ_z^2 and σ_ϵ^2. The intercept term β_0 and the log relative risk, β_1 are assigned vague prior represented by Gaussian distributions with large variances. The details of the MCMC simulations for this model are given in Section 6.3.

8.4.1 Results

Using the measurement error model, the MCMC algorithm were run for 40,000 iterations discarding the first 10,000 as 'burn in'. The estimates of the model parameters can be seen in

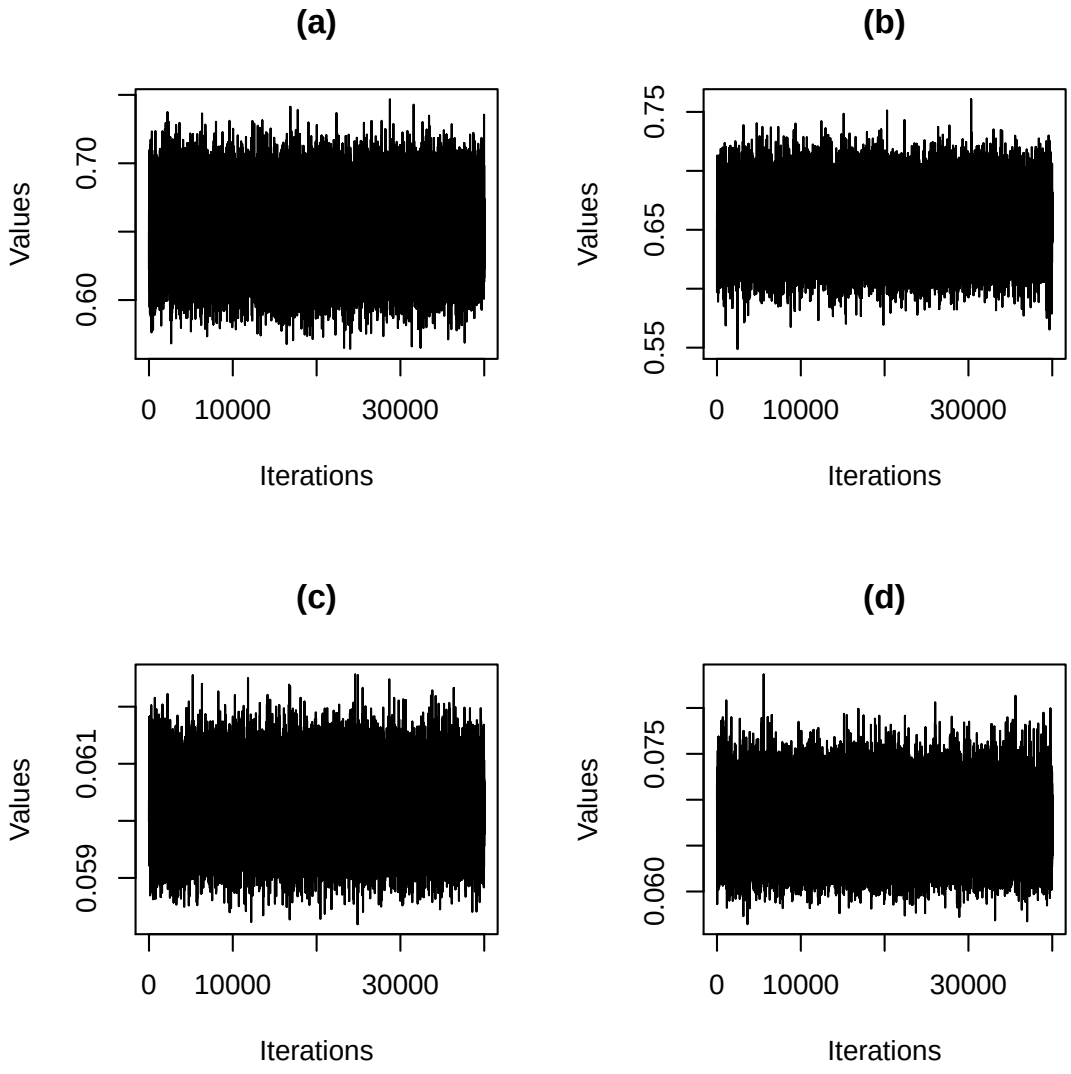

Figure 8-5: The trace plot of sampled posteriors of parameters from the two-stage Bayesian measurement error model implemented by MCMC with 40,000 iterations. (a):ρ, (b):$\mu^{(2)}$, (c):σ^2_ϵ, (d):σ^2_z.

Table 8.3: Estimates of the model parameters for the two-stage model with a measurement error model for the exposures, together with their 95% confidence/credible intervals.

Parameters	Estimated	95% CI
$exp(\beta_1)$	1.0321	1.0137 - 1.0506
β_0	2.9184	2.8791 - 2.9576
ρ	0.6515	0.6055 - 0.6975
$\mu^{(2)}$	0.6571	0.6134 - 0.7009
σ^2_ϵ	0.0603	0.0592 - 0.0613
σ^2_z	0.0675	0.0615 - 0.0735

Table 8.3. The estimated log scaled relative risk parameter, β_1, is 0.0316, meaning the estimated relative risk is $exp(0.0316) = 1.0321$ (95% confidence interval 1.0137-1.0506). Figure 8-5 presents the trace plots of samples drawn from the posterior distributions of parameters. This figure shows that the chains for all the posterior distributions of the parameters are relatively stable and seem to have converged. As the estimates of risk that are produced are very similar to what this may be seen as one of the advantages of the measurement error model over the more complex spatio–temporal model as there are fewer issues associated with the MCMC in the absence of the spatial parameters.

As with using the the spatio–temporal, sensitivity analyses to the choice of priors was performed and in this case, in the absence of the spatial parameters, there was very little difference in the results obtained.

8.5 Conclusion

In this chapter we analysed the air pollution and mortality data from Greater London for 3 consecutive years using three models: standard model, spatial temporal model and measurement error model. For the estimated relative risk, we can see the value conducted from measurement error model is very close to the one from standard model, but the estimated relative risk from spatial–temporal model is obviously larger.

We notice that the estimated correlation–distance parameter from spatial temporal model is 0.0904, this indicates there is very low spatial effect in the collected London data. In addition, the sensitivity analysis shows the posterior distribution of correlation–distance parameter highly depends on the choice of bounds in the prior, and changing the upper bound of the prior leads to higher estimates of the parameter. This suggested that the value of the parameter might

be even greater than estimated, with large values of this parameter indicating very small correlation over even small distances. If it is the case that there is not really any spatial structure in the data then fitting a complex spatial model would be unnecessary and in fact might even introduce bias.

From Table 8.2 and Table 8.3, we can see the estimated measurement error variance is reduced to a half when applying spatial temporal model compared to the value estimated from measurement error model. This suggests that using spatial temporal model is able to to classify the uncertainty in the data, and is very useful when predicting the values from the model. The London data used in this analysis only includes 23 monitoring sties which is a relatively small dataset, hence it may not be able to provide enough information for the parameter estimation in Bayesian settings, we would like to assess the efficacy of spatial–temporal models in a much bigger dataset with more spatial effect in it.

Chapter 9

Incorporating high–dimensional exposure modelling into studies of air pollution and health

The majority of the studies examining the associations between in air pollution and health use data from monitoring networks as a proxy for the exposure to air pollution experienced by the populations in question. Information on ambient concentrations therefore often comes from a number of monitoring sites, often located within an urban area, each of which may measure different pollutants, and in addition may be subject to measurement error and contain periods of missing data.

The amount of monitored air pollution data that is available is increasing dramatically; in London for example there are now more than 80 monitoring sites measuring particulate matter on a hourly, and sub–hourly, basis compared with about 10 in the early nineties. Globally the WHO database on air pollution currently comprises ground-level measurements from 1,600 cities. Being able to utlise the increasing amounts of available data will lead to a more accurate assessments of levels of pollution and more realistic models through increased ability to investigate spatial–temporal dependencies.

Black smoke for example has been measured in the UK since the early 1960s as part of the UK Smoke and Sulphur Dioxide network and its predecessor, the National Survey. The monitoring network, which measures both SO_2 and black smoke, was established in the early 1960s, and by 1971 included over 1200 sites. As levels of black smoke and SO_2 pollution have declined, the network has been progressively rationalised and reduced and by 2006 comprised approximately 100 sites. Over time, many sites have been moved or replaced in order to reflect

changing patterns and levels of pollution, and to reduce redundancy in the network. Therefore there is a large number of missing values in the dataset.

In order to perform health analysis using such data, there will be a need for accurate estimates of air pollution during periods and in locations where there is missing data, either by design (where a monitor is not located or in operation) or due to shorter periods where measurements are not available.

A hierarchical modelling approach provides a natural way of modelling data with complex forms of dependence and the models presented here for that purpose are naturally set within a Bayesian framework. Modelling the entire spatial-temporal structure of an environmental field has in the past often been impractical due to the availability, or lack thereof, of data in the quantities required to produce reliable estimates. Where such data is available, its high dimensionality has meant that the computation required may be prohibitive. There is therefore a need for efficient methods of estimation, particularly in reference to the computational issues that are likely to arise when attempting to implement the models using Markov Chain Monte Carlo (MCMC) sampling. This has led to the development of alternative methods based on approximations within the Bayesian inferential framework and here we specifically consider those based on Integrated Nested Laplace Approximations (INLA) (Rue et al., 2009).

9.1 Hierarchical modelling

In this chapter we develop an approach for estimating the adverse health effects of environmental hazards. This approach incorporates health data, available in aggregate form, and exposure data measured at point locations over space and time. We particularly consider cases where spatio–temporal data records are of such high dimension that conventional computational approaches fail. Before embarking on the description of the health and exposure models, we now briefly review the general framework of the hierarchical Bayesian models described in Chapter 4.

Hierarchical Bayesian models are an extremely useful and flexible framework in which to model complex relationships and dependencies in data. There are three parts to the hierarchy;

- The observation, or measurement, level; $Y|Z, X, \theta_1$. The data is assumed to arise from an underlying process Z which is unobservable but from which measurements can be taken, with error, at locations in space and time. Measurements may also be available for covariates, X.

- The underlying process level; $Z|\theta_2$. This drives the measurements seen at the observation level. It may be, for example, a spatio–temporal process representing an environmental hazard.

- The parameter level; $\theta = (\theta_1, \theta_2)$. This contains models for all of the paramaters in the observation and process level and may control things such as the variabilty and strength of any spatio–temporal relationships.

Here the notation, $Y|X$ means that the distribution of Y is conditional on X. The underlying spatio–temporal process, Z may be viewed as lying in continuous domains of time and space, $\mathcal{T} \subset \mathcal{R}^1$ and $\mathcal{S} \subset \mathcal{R}^d$ respectively, where $\mathcal{R}^d$ denotes d dimensional Euclidean space. However, even when Z is continuously monitored over time, monitors may only report results at discrete times i.e. $\mathcal{T} = \{0, 1, \ldots, N_T\}$ for some N_T. The same may be true over space, where the locations at which air quality monitors can actually be placed may be restricted to a relatively small number of locations, for example on public land, leading to a discrete $\mathcal{S}$ in practice.

The approach developed in this paper involves models for both health counts and exposures and each of these can be framed in the context of a hierarchical model. To avoid ambiguity between the two, we use $Y^{(1)}, X^{(1)}, Z^{(1)}, \theta^{(1)}$ for the health models and $Y^{(2)}, X^{(2)}, Z^{(2)}, \theta^{(2)}$ for the exposure models. It is noted that, although the health counts, $Y^{(1)}$ can be considerd to be measurements from an underlying true level with differences occurring, for example due to missclassification or data anomilies, here we consider them to be an accurate reflection of the truth, i.e. $Y^{(1)} = Z^{(1)}$.

9.2 Integrated Nested Laplace Approximation

Integrated Nested Laplace Approximation (INLA) is recently introduced to perform fast Bayesian inference. It obtains the inference results by accurate numerical approximation of the marginal posterior densities of interested variables and heyperparameters rather than sampling as MCMC, and it provides precise estimate in relatively short time. More details can be found in Rue et al. (2009) and Lindgren et al. (2011).

9.2.1 Latent Gaussian models

Gaussian fields (GF) play an essential rule in spatial statistical modelling. For clarity of explanation, we drop the $Y^{()}, Z^{()}, \theta^{()}$ notation in what follows, and only the spatial domain is considered. For any coordinate s, $Z = (Z_1, \ldots, Z_s)$ is a GF if all sub-collections are joint-

ly Gaussian distributed. The Latent Gaussian model is a latent field in Gaussian distribution regressed to response variables, such as:

- Observation model; $Y_s|Z_s \sim \pi(Y_s|Z_s, \theta_1)$.
- Process model; $Z|\theta_2 \sim N(\mu, \Sigma_{\theta_2})$.
- Parameter model; $\theta = (\theta_1, \theta_2) \sim \pi(\theta)$.

Therefore $\pi(Z, \theta|Y) \propto \pi(\theta)\pi(Z|\theta)\prod_s^{N_S} \pi(Y_s|Z_s, \theta)$. There are two conditions of latent Gaussian model required so that produce fast inference, the first one is that the latent field is normally in large dimension, for instance $10^2 - 10^5$, and it admits conditional independence properties. Secondly, number of hyerparameters is relatively small, for instance smaller or equals to 6. Although GF is widely used in environmental studies, the computational issues have always been a drawback. This is because of the $\mathcal{O}(s^3)$ cost of factorising dense $s \times s$ matrices, beside, the use of hierarchical Bayesian models which needs repeated computations leads to infeasibility.

9.2.2 Gaussian Markov Random Field

A GMRF is a set of Gaussian random variables with Markov properties. The advantage of transposing a Gaussian field to a Gaussian Markov random field (GMRF) is the introduction of GMRF's sparse matrices to numerical methods, hence it is creditable to use GF for modeling but to proceed computation by GMRFs. The Markov property is determined by neighbourhood structure, that contributes to the dependence of full conditional distribution of unit Z_i only base on a few components around it (denoted Z_{δ_i}), hence the set of neighbor of Z_i is as following

$$\pi(Z_i|Z_{-i}) = \pi(Z_i|Z_{\delta_i})$$

where the notation Z_{-i} denotes all elements in Z except Z_i. In this case the precision matrix element $Q_{ij} = 0$ if and only if Z_i and Z_j are independent conditional on Z_{-ij},

$$Z_i \perp Z_j | Z_{-\{i,j\}} \Longleftrightarrow Q_{ij} = 0$$

Hence for a GMRF, the precision matrix Q will be sparse, allowing efficient computation.

9.2.3 INLA inference

INLA uses a Laplace approximation to the posterior distribution of the parameters, θ, given measurements of the response, Y. The aim is to obtain posterior marginal quantities such as $\pi(\theta_i|Y)$ and $\pi(Z_s|Y)$ where for example $\pi(\theta_i|Y) = \int \pi(\theta|Y)d\theta_{-i}$ and $\pi(Z_s|Y) =$

$\int \pi(\theta|Y)\pi(Z_s|\theta,Y)d\theta$. In order to achieve this, approximations need to be built; $\tilde{\pi}(\theta|Y)$ and $\tilde{\pi}(Z_s|\theta,Y)$. The Laplace approximation to the posterior $\tilde{\pi}(\theta|Y)$ is given by

$$\tilde{\pi}(\theta|Y) \propto \left.\frac{\pi(Z,\theta,Y)}{\tilde{\pi}_G(Z|\theta,Y)}\right|_{Z=Z^*(\theta)}$$

where $\tilde{\pi}_G$ is a Gaussian approximation at the mode $z^*(\theta)$ of the conditional distribution of Z given θ. Given such an approximation, numerical integration can be used to evaluate the required integrals. The same procedure can be used to approximate the posterior distribution of $\pi(Z_s|Y)$.

9.2.4 Applications of INLA

Illian et al. (2012) provided a flexible framework for routinely fitting models to complex spatial point pattern data using a model class that accounts for both local and global spatial behaviour. A combination of the flexibility of the log Gaussian Cox process that results from its doubly stochastic structure with the use of constructed covariates is suggested to reflect spatial behaviour. Then INLA is used to fit these models which speeds up parameter estimation substantially such that Cox processes can be fitted within feasible time. Two examples are given in this paper, the first one is the rainforest dataset which is considered as a Cox process model for a point pattern data set with a large number of points and two observed covariates. The second example is the Koala dataset which consists of the locations of 915 eucalyptus trees. These two very different examples indicate that the framework suggested in this paper can be applied in a wide range of situations and is flexible enough to facilitate the fitting of other even more complex models.

Li et al. (2012) analysed the clinical data on the location of residence at the time of diagnosis of new Lupus cases in Toronto, Canada, for the 40 years prior to 2007, and aimed to find areas of abnormally high risk. The inference is complicated because of numerous irregular changes in the census regions on which population is reported, hence they introduced a model consisting of a continuous random spatial surface and fixed effects for time and ages of individuals. The process is modelled on a fine grid and Bayesian inference is performed using Integrated Nested Laplace Approximations. Predicted risk surfaces and posterior excedance probabilities are produced for Lupus and, for comparison, Psoriatic Arthritis data from the same clinic. They concluded that the area in the vicinity of the Lupus clinic is the region of Toronto where spatially varying social or environmental factors could be causing higher incidence of Lupus than would be expected given the population. The price paid for the speed and robustness of INLA has been the need to ignore uncertainty in the age and time effects.

In Eidsvik et al. (2013), they combined computational ideas for modeling and inference of spatial data, and addressed the computational challenges in modeling large spatial datasets by merging predictive process models and INLA. First, they used the predictive process model as a reduced-rank spatial process, to diminish the dimensionality of the model. Then they proceeded to develop a computational framework for estimating predictive process models using the integrated nested Laplace approximation. Results are presented for synthetic data, an environmental dataset and for a large dataset on forest biomass. The predictive process models and approximate Bayesian inference using INLA provide very fast analysis of large spatial data sets.

More recently, Schrödle and Held (2011) focused on the usage of the INLA method for approximate Bayesian inference in parameter-driven models, which can be conducted using a toolbox for generalized dynamic models. They studied the networks of moving individuals such as traded animals between farms, representing a potential risk for the spatio-temporal spread of an infectious disease. Two frameworks of parameter and observation-driven models are proposed to assess the relationship, both of which are discussed in the context of univariate and multivariate time series of counts with specific emphasis on the direct inclusion of network data. In contrast to observation-driven models, where previous cases are included directly, the disease incidence in a parameter-driven model is governed by a latent stochastic process. Ready-to-use software based on INLA is presented for inference in parameter-driven models. In this context, the predictive performance of both formulations is assessed using proper scoring rules and a score regression approach. The impact of cattle trade on the spatio-temporal spread of Coxiellosis in Swiss cows, 2004-2009, is finally investigated.

Cameletti et al. (2011) employed the SPDE approach for a hierarchical spatio-temporal models for particular matter concentration in the North-Italian region Piemonte over the period of October 2005 - March 2006 winter season. The model involves a Gaussian field, affected by a measurement error, and a state process characterized by an AR(1) dynamic and spatially correlated innovations. They used R-library INLA on a Intel Xeon 12 CPU machine to get the parameter posterior estimates together with prediction and uncertainty map, and it took a total time of about 100 seconds for getting the posterior distributions of the hyperparameters and of the latent field over the triangulated domain. As a conclusion, they outlined that the computational strength of the SPDE approach implemented by INLA stands out clearly compared to other machine settings and employed software, an in addition, problems of convergence and mixing typical of the sampling are not an issue at all when working with INLA.

9.3 Health effects models

In order to assess the effect of air pollution on health, models are required that relate risk to exposure, both in terms of the degree of exposure and the time over which the exposure occurred. In cohort studies of individuals, such models need to account for the duration of exposure, time since first exposure, time since exposure ceased and the age at first exposure (Breslow and Day, 1980) (Waternaux et al., 1989). For the development of carcinogenesis, complex multistage models have been developed that use well defined dose-response relationships (Dewanji et al., 1999). However, when using aggregated daily mortality counts for a specific period, e.g. day or health period, and specified area, detailed exposure histories and other information are generally not available.

Considering a generic area for ease of illustration, let $Y_t^{(1)}$ be the health outcome at time t, e.g. the number of respiratory deaths on a single day or other period of time, and the true exposure history $Z_u^{(2)}, 0 \leq u \leq t$, then the outcome is modelled as a function of the exposure history.

$$E(Y_t^{(1)}) = f(Z_u^{(2)}; \quad 0 \leq u \leq t) \tag{9.1}$$

As true lifetime personal exposure to air pollutants is unmeasurable, as they depend on ambient levels and integrated time–activity, the term 'exposure' here relates to cumulative ambient outdoor concentrations of air pollutants, measured at the aggregate area level. The summaries of the exposure history are therefore constructed based on available data, $Y_t^{(2)}$.

If it is assumed that $Z_u^{(2)}$ is piecewise continuous, then the cumulative exposure up to and including time t is

$$\int_{u=0}^{t} Z_u^{(2)} du \tag{9.2}$$

Rather than just considering the effect of the total exposure over a period of time, the contributions from intervals within the period may be of interest, in which case Equation (9.2) can be expressed in the form of weighted integrals (Breslow et al., 1983; Bandeen-Roche et al., 1999).

$$C_t = \int_{u=0}^{t} W_{t-u} Z_u^{(2)} du \tag{9.3}$$

where the weights, W_{t-u}, determine the aspect of the exposure being summarized. For example if the weights are of the form $W(u) = \min(1, u/b)$, then the exposures are phased in linearly over a period of length b until reaching their maximum. This can allow for delayed as well as cumulative effects depending on the form of the weights. In individual studies, the form of the cumulative exposure can be explicitly modelled, for example in the case of exposure to asbestos fibres, where the rate of elimination of the fibres from the lungs, λ, may be incorporated and the model takes the form $W(u) = \{1-\exp(-\lambda u)\}/\lambda$ (Berry et al., 1979).

Since significant exposure to air pollution may start later than birth, the lower limit of the integral may not be zero. Instead the sum is likely to be over a specified period of time. If the weights are of the form

$$W(u) = \begin{cases} 1/(b-a+1) & \text{for} \quad a \leq u < b \\ 0 & \text{otherwise} \end{cases} \tag{9.4}$$

then the summary will represent the average for the period $(t-b, t-a]$, $0 \leq a < b \leq t$. For example, when studying the short term affects of air pollution, with daily measurements of health and air pollution, if $a = 0$ and $b = 2$, then $W(t-u)$ would represent a three day mean.

When dealing with health counts, and exposure measurements made at discrete times, the integral in Equation (9.3) can be approximated by a summation over a suitable discretisation.

$$C_t = \sum_{k=0}^{t} W_{t-k} Z_k^{(2)} \tag{9.5}$$

If the probability of disease given cumulative exposure is assumed to be proportional to $\exp(\gamma C_t)$, i.e. a log-linear model in cumulative exposure, then a Poisson model can be used to estimate the weights, W_{t-k} in Equation (9.5). Assuming that $Y_t \sim Poisson(E_t \mu_t)$ where E_t represents the expected number of cases Breslow and Day (1987) then

$$\log \mu_t = \beta_0 + \gamma \sum_{k=0}^{t} W_{t-k} Z_k^{(2)} = \beta_0 + \sum_{k=0}^{t} \beta_{t-k} Z_{t-k}^{(2)} + \sum_{j=1}^{J} \beta_j X_{jt}^{(1)} \tag{9.6}$$

where $X_j^{(1)}$, $j = 1, \ldots J$, are area-level covariates. Hence the parameters, β_{t-k} represent the effect of exposure k time periods ago. Comparison with Equations (9.3) and (9.5) shows that $\beta_{t-k} = \gamma W_{t-k}$. The expected number of deaths will be $E = \sum_{k=1}^{K} N_k r'_k$, where r'_k are the age–gender specific mortality rates for the reference population (usually a country or other

large area) and $N_k; k = 1, \ldots, K$ are the populations in the area of study, in each age–gender group k. It should be noted that these are not the expected number of cases in the sense of statistical expectation, but are what would be expected based on applying national rates of disease to the population structure of the areas being studied.

It is possible to specify the shape of the distributions of the weights, W_{t-k}. For example, Schwartz (2000) describe the use of a distributed lag model (DLM) within aggregate level studies examining the short term effects of air pollution on health where the weights fit a polynomial function (Harvey, 1981). This requires assumptions to be made on the maximum lags that are likely to have an effect and the smoothness of the patterns over lags, which is determined by the polynomial used, but has the advantage of increasing the stability of the individual estimates where there is high collinearity between the explanatory variables (Zanobetti et al., 2000). The required assumptions have been formulated in terms of priors when implementing DLMs within a Bayesian setting (Welty et al., 2009).

There is a strong possibility of over-dispersion in the Poisson model, where the variance is greater than the mean, arising from the presence of unmeasured confounders. These may be operating at the individual level, e.g. smoking, or at the area level, e.g. residual socioeconomic confounding. Over dispersion may also arise because of data anomalies, i.e. errors in the numerators and/or denominators, e.g. due to migration which may make it unreasonable to assume that $Y^{(1)} = Z^{(1)}$. Making no allowance for the extra-Poisson variability that may be present will lead to confidence intervals for the estimates of risk being too narrow and changes in deviances, used to compare models, being too small. An attempt to correct these effects can be made using a quasi-likelihood (McCullagh and Nelder, 1989).

9.4 Exposure Modelling

Missing values will arise both from short–periods in which monitors were not reporting information and from locations and times for which there were no monitoring sites. One approach is to represent the ambient pollution surface with a spatial or spatio-temporal model, and then to estimate the quantities of interest such as estimated exposures when and where measurements were not taken using prediction methods. As described in Section 9.1, the spatial-temporal random field, $Z_{st}, s \in \mathcal{S}, t \in \mathcal{T}$, is a stochastic process over a region and time period. This underlying process is not directly measurable, but realisations of it can be obtained by taking measurements, possibly with error, at a set of known locations in space $S = \{s_1, ..., s_{N_S}\} \in \mathcal{S}$ and time $T = \{t_1, ..., t_{N_T}\} \in \mathcal{T}$. In a purely spatial analysis, repeated observations at a spe-

cific location over time are treated as independent realisations of the underlying process.

As described in Chapter 2 there are three levels to the hierarchy that we consider. The observed data, $Y_{st}^{(2)}, s = 1, ..., N_S, t = 1, ..., N_T$, at the first level of the model are considered conditionally independent given a realization of the underlying process, $Z_{st}^{(2)}$. The second level describes the true underlying process as a combination of a trend (mean), $\mu_{st}^{(2)}$, and a random process, ω_{st}, which has spatial–temporal structure in its covariance. In a Bayesian analysis, the third level of the model assigns prior distributions to the hyperparameters from the previous levels. Thus in summary we have:

$$\begin{aligned} Y_{st}^{(2)} &= Z_{st}^{(2)} + \epsilon_{st} \\ Z_{st}^{(2)} &= \mu_{st}^{(2)} + \omega_{st} \end{aligned} \tag{9.7}$$

where the $\{\epsilon_{st}\}$ are an independent random, or measurement, error terms, $\mu_{st}^{(2)}$ is a space-time mean field (trend) and ω_{st} is a spatial–temporal process.

The second line in Equation (9.7) comprises a mean function together with a zero-mean spatial–temportal process. Previous studies have modelled the mean function with a trend surface model (Zhu et al., 2003), cyclical variation (Tonellato, 2001), a temporal only trend (Zidek et al., 2014; Shaddick and Zidek, 2014) and the Kriged-Kalman model (Sahu and Mardia, 2005). The spatial–temporal process can be considered to be the combination of three components; space, time and space–time interaction. These three components may be combined in either additive or mulitplicative form Sahu and Mardia (2005); Zhu et al. (2003). For the former we have:

$$\omega_{st} = m_s + \gamma_t + \kappa_{st} \tag{9.8}$$

This form has been used by a number of authors to model ambient air pollution, including for example Zidek et al. (2014) , Sahu et al. (2007), Sahu and Mardia (2005) and Sahu et al. (2006) who modelled PM_{10} in Vancouver and $PM_{2.5}$ in Ohio state, New York City and a collection of midwestern U.S. states, respectively. In a *separable* model, the spatial and temporal components are considered entirely separately with no interaction between them, i.e. $\kappa_{st} = 0$.

It is commonly assumed that the spatial effects, m_s, represent a stationary spatial process with the relationship between correlation and distance between the sites being represented by a function from the Matern family of correlation functions, which takes the following form,

$$\frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}(2\sqrt{\nu}t\phi)^{\nu}K_{\nu}\sqrt{\nu}t\phi \tag{9.9}$$

where $K_\nu(\theta \parallel h \parallel)$ is a modified Bessel function of the second kind, σ^2 is the overall variance and (ν, ϕ) are parameters that control the smoothness and strength of the distance–correlation relationship respectively.

9.4.1 Prediction at unsampled locations

The posterior distributions for the underlying level at a point s_0 not included in the sampled locations is

$$\begin{aligned} p(Z_0^{(2)}|Y^{(2)}) \propto p(Z_0^{(2)}, Y^{(2)}) &= \int ... \int p(Y^{(2)}, Z^{(2)}, Z_0^{(2)}) dZ^{(2)} \\ &= \int ... \int \left\{ \prod_{s=1}^{S} p(Y_s^{(2)}|Z_s^{(2)}) \right\} \times \dots \\ \dots &\times p(Z_0^{(2)}|Z^{(2)}) p(Z^{(2)}) dZ^{(2)} \end{aligned}$$

This form can be further expanded to incorporate the conditioning on the parameters within the model, i.e. $p(Z^{(2)}|\psi, \nu, \phi)$, where ψ are the coefficients in the mean term and (ν, ϕ) those in the variogram/covariance function. In this way the uncertainty in the estimation of the parameters of the spatial–temporal model can be 'fed' through to the predictions.

9.4.2 Inference

For Bayesian analyses, the posterior distributions will often involve high dimensional integration and may be analytically intractable. However, samples from these distribution may theoretically be generated in a straightforward fashion using MCMC sampling (Smith and Roberts, 1993). The main constraint for this approach, particularly when using large spatial datasets, is its demanding computational requirements. This can be both because of the need to manipulate large matrices within each simulation of the MCMC and also in the lack of convergence of parameters estimates in complex models (Finley et al., 2007).

The increasing size and complexity of experiments and the databases they generate have outpaced the speed of readily available computational hardware. This has forced the development of of practical alternatives to MCMC algorithms.

Here we concentrate on recently developed techniques which perform approximate Bayesian inference based on integrated nested Laplace approximations (INLA) and thus do not require full MCMC sampling to be performed (Rue et al., 2009). INLA has been developed as a computationally attractive, practical alternative to MCMC.

In a spatial setting the INLA approach provides a natural approach to modelling areal level data. Applying the approach to point level data of the type that will arise from air pollution monitoring sites can be performed by using a link between Gaussian Fields (GF) with Matern covariance functions and Gaussian Markov Random Fields (GMRFs) through use of Stochastic Partial Differential Equations (SPDE) (Lindgren et al., 2011).

Lindgren et al. (2011) show that a field with a Matern covariance structure can be expressed as the solution of an SPDE. If a GF, Z, has a Matern spatial covariance as given by (9.9) then it is the solution of the SPDE

$$(\kappa^2 - \Delta)^{\alpha/2} z_S = \mathcal{W}_S, S \in \mathcal{S}, \alpha = \nu + d/2, \kappa > 0, \nu > 0 \tag{9.10}$$

where $(\kappa^2 - \Delta)^{\alpha/2}$ is a pseudo–difference operator, Δ is the Laplacian and $\mathcal{W}$ is spatial white noise with unit variance.

This SPDE in turn can be approximated using a finite element method, using triangulatin over the spatial domain of interest. An induced GMRF representation of the original GF can then be found with the precision matrix being approximated by a sparse precision matrix, Q. This represents the information within the covariance matrix of the original GF, Σ and its sparsity allows computational efficiency. The GMRF is used by INLA for performing computations that would be computationally prohibitive using the GF directly.

By combining the temporal and spatial components of the model, predictions can be made at locations where there are missing values for certain years.

9.5 Linking exposure and health models

Pollution data are generally obtained from N_S fixed site monitors located within $\mathcal{S}$, each of which will measure ambient pollution concentrations continuously throughout the year.

The set of pollution monitoring sites are collectively denoted by $S = s_1, \ldots, s_{N_S}$, where $s_l = (a_l, b_l) \in \mathcal{R}^2$. However health data are commonly available only at aggregated level for administrative areas, $A_i, i = 1, ..., N_{A_i}$ and therefore a suitable summary of the concentrations in an area for a particular time period is required. The true mean exposure for time t in a health area, A_i is given by

$$Z_{it}^{(2)} = \int_{s \in A_i} N_s Z_{st}^{(2)} ds \tag{9.11}$$

where $Z_{st}^{(2)}$ is the ambient pollution concentration at all possible locations s in A_i at time t and N_s is the population density such that $\int_{s \in A_i} N_s ds = 1$. However the information required to perform the integral will be unavailable. Therefore there is a need to approximate this. As described in Chapter 4, the simplest and most commonly used approach being to take the average of the observed measurements from actual monitoring sites located within the health area,

$$\overline{Y}_{it} = \frac{1}{N_{A_i}} \sum_{s \in A_i} Y_{st} \tag{9.12}$$

where N_{A_i} is the number of monitoring sites located within area A_i. Here missing values are typically ignored, something that can lead to bias if there are strong temporal trends in the data. An example of this can be seen in the case study presented in Section 9.6.

Alternatively, an exposure model can be used to provide the required information including using predictions in place of any missing values. Any approach for using such predictions in the health model must acknowledge the uncertainty in the predictions and allow for it to be incorporated in final measures of uncertainty, and confidence intervals, associated with those measures of risk.

As described in Chapter 5, Section 4.3, we advocate a two-stage approach to modelling the exposures and using predictions in the health model with the associated uncertainty acknowledged using multiple imputation. This allows the uncertainty in predictions to be represented by using a set of plausible values for the exposures which comprise samples from the posterior distributions of the predictions at the required locations in space and time. Taking M multiple (joint) samples from the posteriors results in M multiple datasets which are repeatedly used in the health model.

This requires the ability to draw joint samples from the posterior distributions of the predictions from the exposure model. This is possible in the R-INLA package using the function

`inla.posterior.sample`. In computing the approximation to the required distibutions, $\tilde{\pi}(\theta|\mathbf{y})$ $\tilde{\pi}(z_{ts}|\theta,\mathbf{y})$, R-INLA uses numerical integration based on interpolation between a number of chosen 'integration points' (Rue et al., 2009). Taking $\tilde{\pi}(z_{ts}|\theta,\mathbf{y})$ as an example, the integration points are selected from a set of candidate points on a grid. After exploring $\log(\tilde{\pi}(z_{ts}|\theta,\mathbf{y}))$ to find the mode, a point is selected if the difference between $\log(\tilde{\pi}(z_{ts}|\theta,\mathbf{y}))$ evaluated at that point and the value evaluated at the mode is greater than a prespecified constant. Apart from the integration based on this procedure for finding approximations to the marginal distributions, the information stored about the distribution at these integration points can be kept. This allows the function `inla.posterior.sample` to be used after the main INLA run. Joint samples from the posteriors can be obtained by sampling from Gaussian approximations at the integration points for all of the parameters, including predictions from the exposure model. A combined analysis of these datasets is then performed using multiple imputation (as described in Chapter 5, Section 4.3). This results in valid statistical inferences that properly reflect the uncertainty due to missing values.

9.6 Case study

The UK black smoke and sulphur dioxide network measured black smoke (BS) sulphur dioxide (SO_2) from the early 1960s until 2006. During that time, at its peak it comprised of over 1200 sites (in the early 1970s). As levels of BS and SO_2 declined from the very high levels in the 1960s, the network dramatically reduced in size and by 2005, shortly before it ceased operation, it contained 65 sites. Over this time, there was a marked decline in the concentrations of BS which can be seen in Figures 9-1, which shows the average levels per year and Figure 9-2 which shows measurements for a selection of individual sites from the network. For further details of the long–term changes in levels of BS and changes in the network see Shaddick and Zidek (2014).

Data were obtained for a total of 3016 sites throughout the operation of the network, of which 2137 sites were designated as being located in residential areas. The locations of the monitoring sites were linked using GIS, as described in Elliott et al. (2007), to electoral wards which is the resolution of the health data. The locations of the wards, together with an indication of the average concentrations of BS over the study period can be seen in Figure 9-3. This allows analyses of the association between health and air pollution to be performed, however there will be areas in which health data is available but monitoring information was not available.

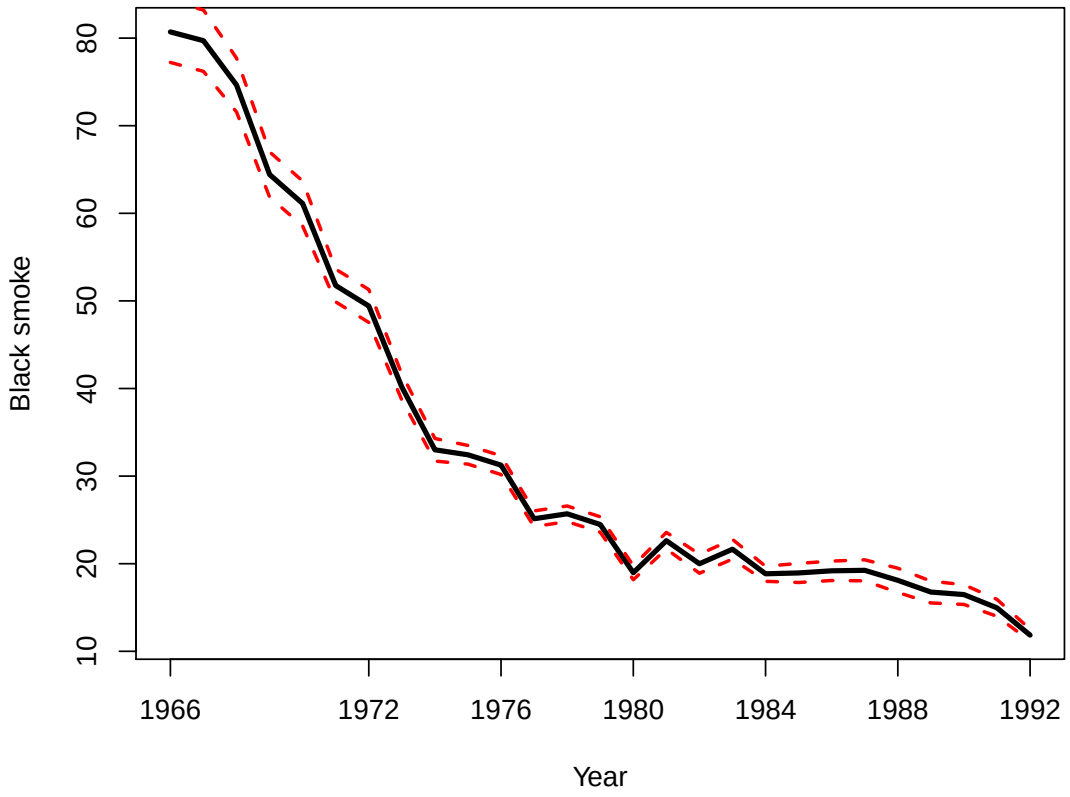

Figure 9-1: Yearly mean concentrations of Black Smoke ($\mu g m^{-3}$) from 1966 to 1992 with associated 95% confidence intervals.

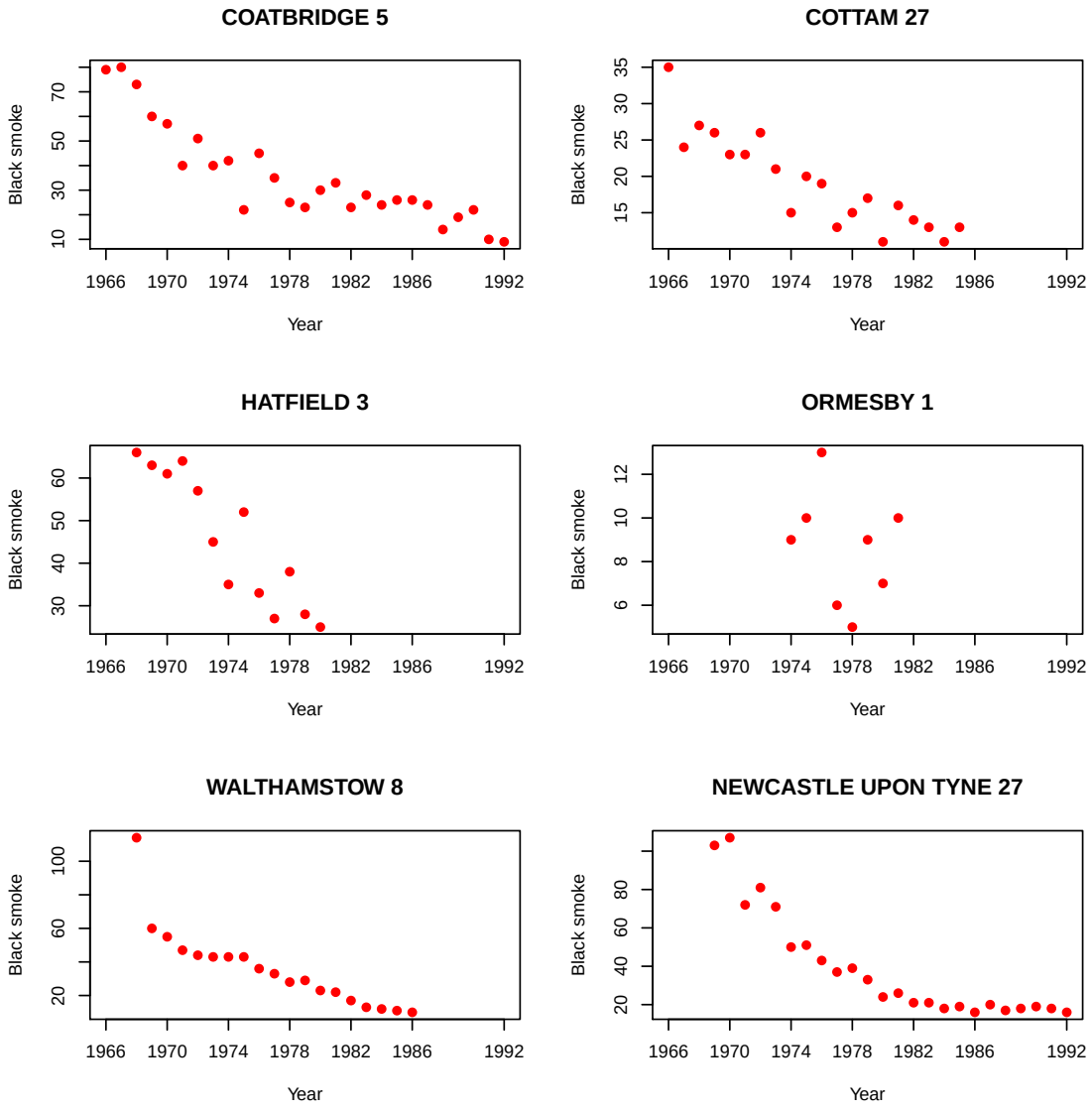

Figure 9-2: Black smoke concentrations ($\mu g m^{-3}$) against time (1966 to 1992) for a selection of sites from the network.

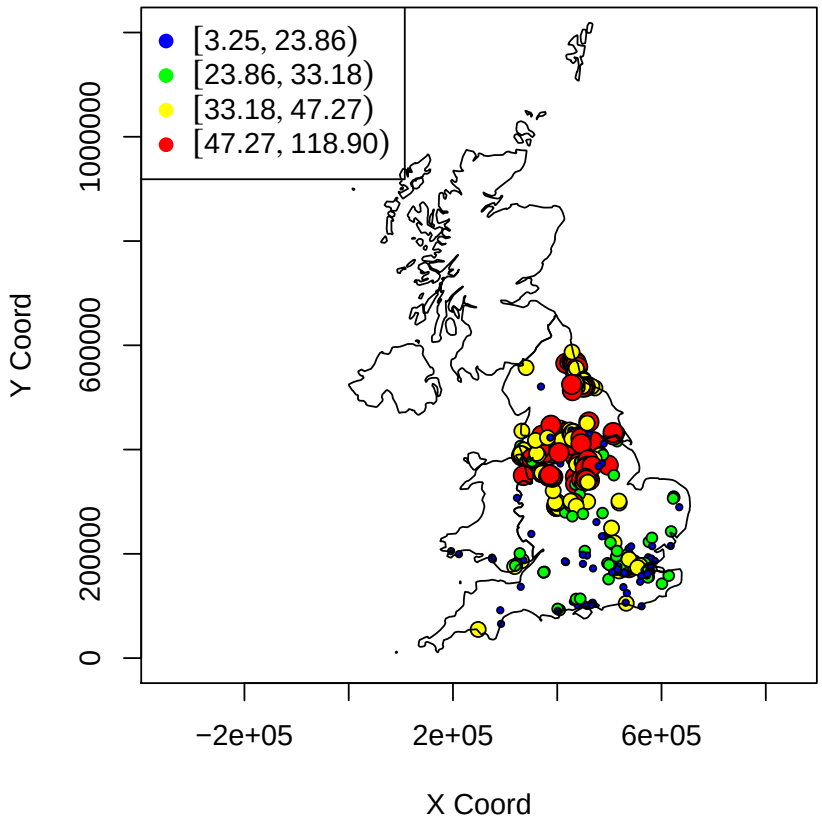

Figure 9-3: Yearly mean concentrations of black smoke ($\mu g m^{-3}$) measured at electoral wards within the UK, 1966-1992.

The health data consist of mortality counts for the period 1993–96 for respiratory diseases in the over 65s. These data were extracted for all ages by ward from national postcoded mortality data, by age and sex, for the period 1993–96. Expected numbers, standardised by age and sex, were calculated for each ward using national mortality counts and population data from the 1991 census. Smoking is known to be a major risk factor for cardio-respiratory illness and it is known that smoking habits vary with social class (Kleinschmidt et al., 1995) and may therefore correlate with pollution levels, and act as a potential confounder. In the absence of data on smoking levels, an area level measure of socio–economic deprivation is used (Carstairs and Morris, 1989), which has previously demonstrated to be related to smoking rates (Kleinschmidt et al., 1995). Summaries of the observed and age-sex standardised expected numbers together with a measure of the relative risk in the health period of 1993-1996 are shown in Table 9.1.

Table 9.1: Summary of respiratory mortality at ward level for 1993–96 at ward level. Means, standard deviations, medians, ranges and interquartile ranges are given for observed (Obs) and age-sex adjusted expected numbers of cases (Exp), together with measures of relative risk (O/E ratio) and the number of wards (N).

	N	Min.	25%	50%	75%	Max.	Mean	SD
Obs	123	3.0	41.5	68.5	107.5	217.0	77.9	46.6
Exp	-	1.9	40.3	63.5	83.3	220.2	64.0	33.9
O/E	-	0.5	1.0	1.1	1.5	2.1	1.2	0.3

The period of study is chosen to represent a time (for the health period) which follows an extended period during which there were great changes in the levels of BS. Studies of the chronic effects of pollution have largely considered concurrent exposures. Over recent decades, air pollution concentrations have generally fallen, in response to industrial and technological changes and more rigorous regulation. At the same time, the character of air pollution has changed markedly, as domestic and industrial coal-burning has declined and emissions from road traffic have increased. Health risks determined on the basis only of current or recent exposures may therefore be misleading, especially for older age groups who may in the past have been subject to very different exposure regimes. Here we use exposures over the previous 27 years.

9.6.1 Statistical modelling

Exposure modelling

Let $Y_{st}^{(2)}$ be the concentration of black smoke measured at location, s, at time, t. Ott (1990) has suggested that a log transformation is appropriate for modelling pollution concentrations, because in addition to the desirable properties of right-skew and non-negativity, there is justification in terms of the physical explanation of atmospheric chemistry. We adopt a similar model to that presented in Shaddick and Zidek (2014) and model the change in levels of BS over time using a random effects model with a quadratic relationship between time and concentrations of BS.

Initial data analysis consisted of two main components following the work of Shaddick and Zidek (2014). This involved looking at temporal trends (as in Figures 9-1 and 9-2) and spatial patterns (as in Figure 9-3) in the data. The rationale for the temporal part of the model was to try to find a form which would both fit the data but also provide a non-complex relationship which might explain the observed decline in concentrations over time. A quadratic effect of time proved a suitable fit to the data and allowed a relatively simple model to be used.

$$Y_{st}^{(2)} = (\beta_0^{(2)} + \beta_{0s}^{(2)}) + (\beta_x^{(2)} + \beta_{xs}^{(2)})t + (\beta_{x^2}^{(2)} + \beta_{x^2s}^{(2)})t^2 + \epsilon_{st} \tag{9.13}$$

where $s = 1, \ldots, N_S$ denotes the site and $t = 1, \ldots, N_T$ the year. The model includes both linear and quadratic effects, $\beta_x^{(2)}$ and $\beta_{x^2}^{(2)}$ of time reflecting the shapes of decline in levels of black smoke observed in the data. The ϵ_{st} is a random error term, which is assumed to be Normally distributed, $\epsilon_{st} \sim N(0, \sigma_\epsilon^2)$. Site specific random effects, $\beta_{xs}^{(2)}$ and $\beta_{x^2s}^{(2)}$ and $\beta_{0s}^{(2)}$, are assigned to the slopes of the linear, quadratic and intercept components respectively. These are constrained to sum to zero, around fixed effects, $\beta_0^{(2)}$, $\beta_x^{(2)}$ and $\beta_{x^2}^{(2)}$ respectively. After allowing for the effects of time, there is likely to be spatial structure in the residuals and therefore the random effects are multivariate normally distributed, $\beta^{(2)} \sim MVN(0, \sigma_s\Sigma)$, with the structure of the covariance reflecting any spatial auto-correlation as in Equation (9.9).

The spatial residuals from the fitted model can be seen in Figure 9-4 which does not appear to show any serious spatial patterns, which is to be desired as the aim is to model the spatial structure within the model and not leave it in the residuals.

It is noted that the model allows for (spatial) random effects on both the intercept and slope terms. Using a simpler model, with random intercepts but fixed slopes turned out to have an interesting side effect which can be seen when comparing maps of the spatial effects from models with and without this flexibility for the slope terms. This can be seen by comparing

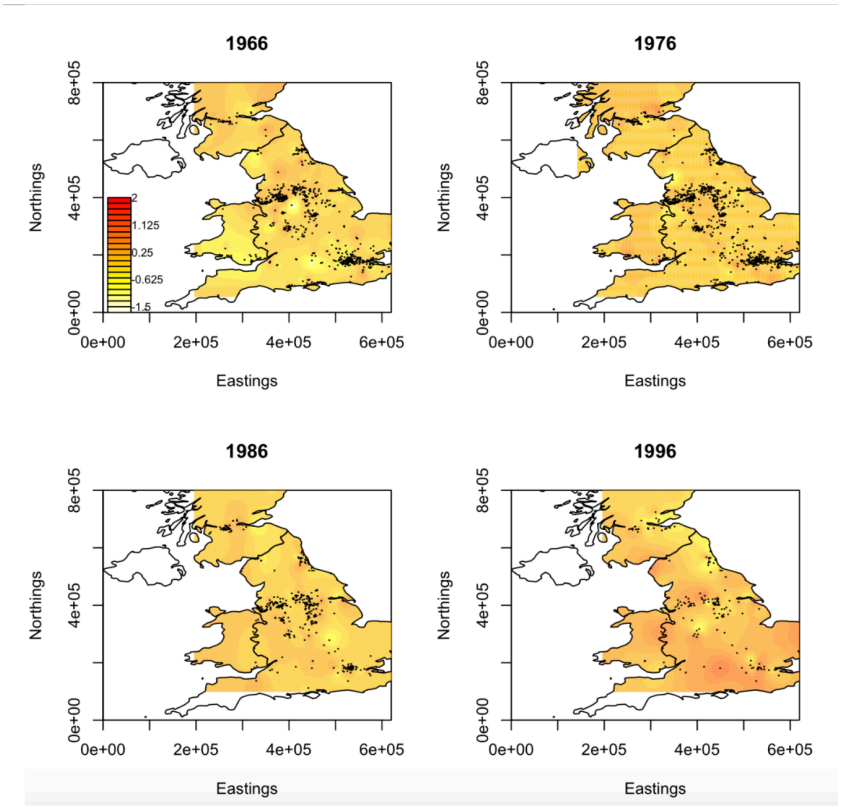

Figure 9-4: Residuals from the spatialCtemporal model for black smoke concentrations (on log scale) by decade.

Figures 9-5 and 9-6 which show maps of the spatial effects from models with fixed and random slopes respectively. In the first case, there is much less spatial smoothing reflecting the fact that fitting a fixed slope is essentially just subtracting a mean term (over time) with the random effects for the intercepts just reflecting the concentrations at individual sites at the beginning of the study period. In contrast, the map shown in Figure 9-6 shows clear smoothing over space and indicates that it would be much more suitable for predictions at times for which there was no recorded measurements at a particular time.

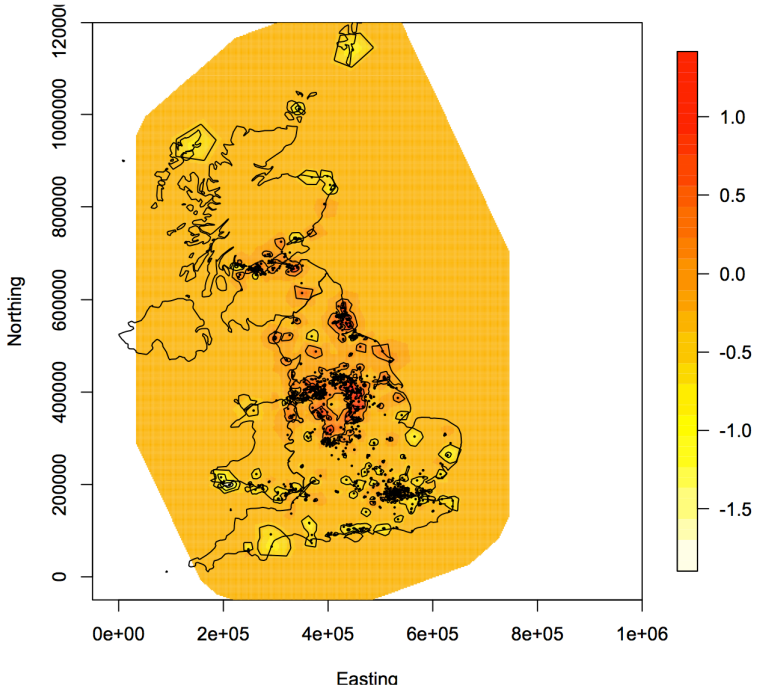

Figure 9-5: Map of the means of posterior predicted distributions black smoke concentrations on the logarithmic scale from a model with spatial structure on the intercepts but with fixed slopes.

Health modelling

Expanding Equation (9.6), we model the number of counts in area i for time t (defined as 1993–96 for this analysis rather than a single year) as Poisson, $Y_{it}^{(1)} \sim P(E_i\mu_{it})$ where E_i represents the expected number of cases in area i for the period from which the health data arise. The log of the rate, μ_{it} is modelled as a function of the levels of air pollution over the previous 27 years with the area level covariate, $X_2^{(1)}$ representing deprivation. As in Equation

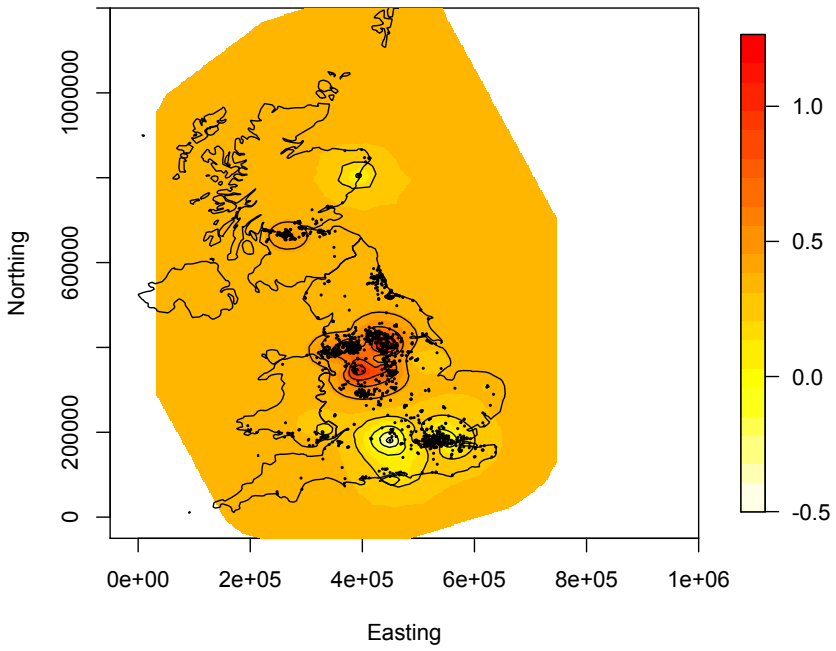

Figure 9-6: Map of the means of posterior predicted distributions on the logarithmic scale from a model with spatial structure on both the intercepts and slopes

(9.6) we use equal weights for each year and use the average pollution over the chosen time period for each area.

$$\log \mu_{it} = \beta_0^{(1)} + \beta_1^{(1)} \tilde{Z}_{s_i} + \beta_2^{(1)} X_{2i} \tag{9.14}$$

where $X_2^{(1)}$ is the area-level index of deprivation with associated coefficient $\beta_2^{(1)}$ and $\beta_1^{(1)}$ which represent the effect of the previous 27 years of exposure which for area A_i is represented by $\tilde{Z}_i$.

Estimates of the exposure for each area, $\tilde{Z}_i$ can be obtained in a number of ways and here we consider three methods.

M1: The average of the available data.

M2: The average of predictions from the spatial–temporal exposure model.

M3: A combination of available data and predictions from the exposure model, enabling missing data to be 'filled in'.

In both M2 and M3 there will be 27 values used in calculating the average over the previous 27 years. However, in M1 the fact that data may be missing is ignored and takes no account of the fact that in some cases the average may be based on small numbers. When there are clear trends in the data, as there are here, the times at which data are available may strongly affect the resulting summary of exposure. For example, if levels are decreasing then missing data at the beginning of the period will result in an underestimate of the overall exposure as higher values will be excluded. Similarly, missing data in the later period when exposures are lower would result in an overestimate.

For methods M2 and M3, multiple imputation is performed by drawing samples from the posterior distributions of the predictions in order to acknowledge the uncertainty that is associated with predicting from the exposure model. One hundred sets of data were produced, comprised of either a combination of available data and predictions (M3) or just predictions (M2). In order to allow for the possibility of extra–Poisson variation, we perform all analysis using both Poisson and quasi–likelihood models.

9.7 Results

Table 9.2 shows the estimated relative risks per $10\mu\text{gm}^{-3}$ of black smoke together with their corresponding 95% confidence intervals obtained from applying the three approaches described

in Section 9.6.1. For each approach, relative risks are estimated with and without adjustment for deprivation using two models; likelihood based Poisson and quasi-likelihood. Results for methods M2 and M3 are obtained from multiple imputation of 100 samples from the joint posterior distribution of the exposure predictions.

Table 9.2: Relative risks (RR) of respiratory mortality, with 95% confidence intervals for an increase of 10 ppb of BS over the previous 27 years. Exposure values are obtained using three methods:(1) using observed data; (2) using predictions from a spatio–temporal model; (3) using observed data combined with predictions to fill in missing values. Risks are estimated with and without adjustment for deprivation using two models; likelihood based Poisson and quasi-likelihood. Results for methods 2 and 3 are from multiple imputation using 100 datasets (see text for details).

Method 1: Observed exposures only

	Without deprivation		**With deprivation**	
	RR	95% CI	RR	95% CI
Poisson	1.037	1.025–1.050	1.038	1.023–1.049
Quasi	1.037	1.005–1.071	1.036	1.003–1.071

Method 2: Predictions

	Without deprivation		**With deprivation**	
	RR	95% CI	RR	95% CI
Poisson	1.022	1.014–1.030	1.021	1.013–1.029
Quasi	1.022	1.002–1.042	1.021	1.001–1.042

Method 3: Observed data and predictions combined

	Without deprivation		**With deprivation**	
	RR	95% CI	RR	95% CI
Poisson	1.011	1.004–1.018	1.010	1.003–1.017
Quasi	1.011	0.994–1.028	1.010	0.992–1.028

For the first approach, based on the given data, there is a significant increase in risk associated with increased levels of black smoke when using the Poisson model (RR=1.037, 95% CI; 1.025 , 1.050) and this result remains significant when using the quasi–likelihood, despite the wider confidence interval; (1.005, 1.071). Significant increases in risk are also seen after adjustment for deprivation. Little difference was observed when adjusting for the effects of deprivation. Although this measure of deprivation has been used in many small-area epidemiological studies (Elliott et al., 1996; Dolk et al., 1997; Elliott et al., 2013) and has been shown to provide a good measure with which to discriminate between poor health associated with deprivation and vice–versa, the score is defined on a national level. To a great extent, the areas studied here, i.e. those that have air pollution monitoring sites located within them, constitute a

set of deprived areas. Deprived areas are likely to have higher levels of pollution (Elliott et al., 2007) and thus, due to the practice of locating monitors in locations where pollution might be expected to be highest, these areas are also more likely to be the ones in which monitoring sites are located. In fact, over 70% of the areas in this study lie in the two most deprived quintiles (over the UK), which would greatly reduce the discriminatory power, in that there would be little to differentiate between a large number of the wards as they would be assigned similar (high) scores.

As discussed in Section 9.6.1 there is the strong possibility that the results based solely on the available exposure data will be biased if there are strong temporal trends, as in this case where there is a marked decline over time. The availability of the exposure data (at ward level) can be seen in Figure 9-7 which shows the years for which information was available over the period 1966-1992.

Using approach M2, predictions from the exposure model are used, not just to fill in missing values in the data for times/locations where data was not measured, but also to replace the measurements where they were available. In doing this, as with any model of this type, very high and low values of the exposures will be smoothed towards the mean. However it is precisely the high values that are likely to be driving the health risk and the combination of these high values together with the low ones which will provide the contrast, i.e. the range of values, that is so important for estimation in any regression model. Using this approach, increased risks are observed for both the Poisson model (RR=1.022, 95% CI; 1.014-1.030) and the quasi–likelihood (95% CI; 1.002-1.042), although the increase is smaller than that observed when using approach M1. The risks again remain after adjustment for deprivation.

Approach M3 uses a combination of the available data with predictions from the exposure model used when measurements are not available. As such, as much as is possible it retains the contrasts in the exposures (unlike approach M2), while having a 'full' set of data over time for each area which will reduce the effect of the bias seen in approach M1. In this case the estimated relative risk is RR=1.011 (95% CI; 1.004 - 1.018) for the Poisson model, without adjustment for deprivation, with the lower end of the CI being just less than one when using the quasi–likelihood model.

9.8 Conclusion

In this chapter we have incorporated large-scale modelling of air pollution over space and time into epidemiological analyses. In performing epidemiological analyses of the relationship be-

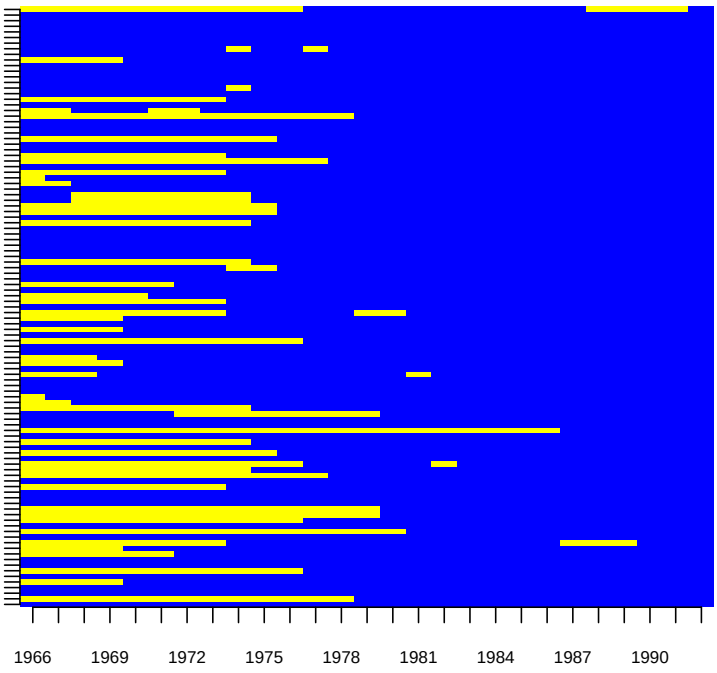

Figure 9-7: Schematic showing the years for which monitoring sites were operational and those when they were not during the period of exposure; 1966-1992. Data are aggregated to the health area (ward) level. Each line represents a ward, with yellow lines showing times where there were no operational monitoring sites and blue lines where monitoring sites were operational and data available for analysis.

tween environmental hazards and adverse health outcomes often there will be locations and periods of time in which exposure information will not be available. This may be due to a fault in monitoring equipment or may be due to the design of monitoring networks and changes over time. In such cases, a direct comparison of the exposure and health outcome is often not possible without an underlying model to align the two in the spatial and temporal domains.

For the Bayesian framework used in this chapter, two stage approaches separate the exposure and health components, whilst still allowing uncertainty from the exposure modelling to be incorporated into the health model. We use multiple imputation based on samples from the joint distribution of the posterior distributions for predictions of the exposures. In approaches M2 and M3 the width of the confidence interval associated with the estimate of risk will incorporate both the uncertainty associated in the estimation of the risk parameter within each of the datasets and also that between the datasets; the latter reflects the uncertainty from the estimation of the exposures.

In the case study, we have attempted to isolate monitoring sites that might indicate the exposures experienced by the populations at risk by selecting only sites that were designated to be in residential areas. However there is the strong possibility that monitoring sites will have been located in areas that were expected to have high concentrations, as may be the case when assessing whether guidelines and policies are being adhered to. This leads to preferential sampling, where when the process that determines the locations of the monitoring sites and the process being modelled (concentrations) are in some ways dependent. Zidek et al. (2014) showed that there is a significant association between measured levels and the probability of a site remaining in the network. They also presented a method for adjusting summary measures (of the levels of pollution) for changes in the monitoring network and preferential sampling. Future research topics may include the possibility of incorporating adjustments directly into the estimation of health risks.

In theory, it would be relatively straightforward to fit the models considered here using MCMC and this would provide a natural way of allowing the uncertainty associated with using predictions from the exposure model to be fed through to the estimates of the health risks. However, in practice the computational requirements may prove to be prohibitive, both because of the requirement to manipulate large matrixes within each simulation of the MCMC and also in convergence of parameters in complex models. Convergence of the spatial parameters in particular can cause problems, especially if datasets are relatively small in which case there might not be enough information to estimate them accurately.

The models considered here were fit using INLA to allow point referenced spatial components to be incorporated. In terms of prediction at a very high number of locations, techniques such as INLA, which perform 'approximate' Bayesian inference and thus do not require full MCMC sampling, provide an extremely appealing approach. In many cases, the underlying field will not be stationary. Bornn et al. (2012) showed evidence of non-stationarity in black smoke concentrations and this is likely to occur with air pollution where many factors, such as topography and wind patterns will affect local concentrations. INLA can be extended to cover non-stationary random Gaussian fields and future work will involve integrating predictions from non-stationary exposure models into health models. Overall, the implementation of the INLA approach in this paper demonstrate how the methods can provide a remarkably fast computational algorithm for application over large domains when standard computational methods might fail.

Chapter 10

The effects of preferential sampling in environmental health effects analyses

In this chapter we assess the potential effects of preferential sampling to the estimation of health risks associated with air pollution, a potentially very important subject that has not received much attention. We start with some background on the subject of preferential sampling. Then, simulation studies that demonstrate the effect it may have on the estimation of the health risks associated with air pollution. In Section 10.2, we then describe an approach to enable the assessment of the potential effects of preferential sampling based on changes in a network over time. We then apply this to a case study of the effects of black smoke on health in the UK during a period in which levels of air pollution were dramatically decreasing and the potential for preferential sampling has been shown to be high.

10.1 Preferential sampling

Preferential sampling is a common phenomenon in environmental studies, as the monitoring locations in a spatial network are often chosen based on a subjective purpose, such as the change of government policies and the intention of monitoring high levels of pollution. For example, if monitors are positioned close to known pollution sources, such as at the roadside near an industrial polluter, or within a city center, then the estimated pollution surface is likely to be overestimated. Both the number and locations of the pollution monitors will affect the accuracy of estimating the true exposure surface. It is often intrinsically assumed that the true exposure surface is based on the random sampling of the complete temporal–spatial pollution field. However, this is extremely unlikely to be the case and the exposure measurements obtained from preferentially sampled networks may lead to an inaccurate estimation of exposure to air pollution and consequently to the estimation of relative risks in epidemiological studies.

Recently there have been a small number of papers published on the subject of preferential sampling in an environmental setting, which occurs when the process that determines the locations of the monitoring sites and the process being modelled (air pollution concentrations) are in some ways dependent. Diggle et al. (2010) extend the classical geostatistical model in two ways; (i), the monitoring locations are treated as random quantities of a log-Gaussian Cox process rather than being fixed; (ii) the exposures are modelled conditionally on the locations assuming a Gaussian spatial process. Through simulation examples they show that ignoring preferential sampling can lead to misleading inferences, especially with spatial predictions. Pati et al. (2011) adapt this approach within a Bayesian framework and demonstrate its use in a case study of ozone data over eastern U.S.A which shows significant evidence of preferential sampling. Other examples of the application of this approach include Lee et al. (2011) who implement it when constructing air quality indicators for a case study set in Greater London.

Gelfand et al. (2012) suggested another approach to dealing with the effects of preferential sampling. Again, the locations are also treated as a realisation of a random process but they use a deterministic model with informative covariates, such as population density, to indicate the underlying pollution surface. This approach is based on the assumption that if sampling locations are drawn as a reflection of covariate factors, then the covariates should be used in the exposure model to correct the preferential sampling bias. A simulation study shows the spatial predictions of exposures under preferential sampling are substantially biased when compared to those from random sampling. Lee and Shaddick (2010) investigated the influence of preferential sampling to the pollution concentration estimation on spatial prediction using a Bayesian spatio–temporal model, again showing significant biases in spatial predictions.

The majority of research in this area has focused on the predictions of exposure surface in a spatial network. The approach in Diggle et al. (2010) models the spatially continuous unobserved process to be used in the stochastic model of locations, but this process is unknown in practice, so it is difficult to specify its propriety. In addition, only a single realization of the underlying random field is used to generate the data for the locations in the study region. For the approach proposed by Gelfand et al. (2005), the difficulty is that it requires the complete information of covariates used in the deterministic model, which is normally unavailable in practice.

An alternative approach to adjusting for preferential sampling from the spatial modelling approaches described above is that of response biased regression modelling (Scott and Wild, 2001). Zidek et al. (2014) proposed a new method to model preferential sampling in environmental networks based on this approach. The idea is based on concepts from survey sampling

in which sampling weights define the under- or over-sampling of specific demographic groups. Resulting estimates can then be adjusted using the sampling weights to allow for the non-random design. They used the Horwitz-Thomson (HT) estimator to unbias estimates based on preferentially sampled data. In short, the HT estimator weighs each observation against the probability that the particular observation is included in the sample. In the setting considered here however the sampling weights, which define the process of preferential sampling, are generally not known. The selection probabilities cannot therefore be characterised as they are in survey sampling. The idea of Zidek et al. (2014) was to estimate these probabilities using logistic regression based on concentrations measured in previous years and locations.

10.1.1 Simulation study

We now investigate the possible impacts of preferential sampling on the estimation of health effects using a series of simulation studies. For simplicity, the simulations represent the generation of data representing a single time period as the goal is to show the effect of preferential sampling in estimating health risks rather than estimating exposure surfaces. The overall aim is to compare the results from the models using the entire data available with those using data that has been intentionally preferentially sampled. Where possible, the parameters used in generating the simulated data are informed by the data used in the case study in Section 10.3.

Data generation

The study region, $\mathcal{S}$, is a unit square 10×10 lattice comprising 100 spatial cells A_i ($i = 1, \ldots, 100$). Within each area, twenty values of the exposure in each grid are generated from a normal distribution in order to represent the exposure collected at $N_{s_i} = 20$ monitoring sites. A demonstration of the lattice is shown in Figure 10-1. The mean values of the distribution within each area are drawn from a uniform distribution, $U(20, 60)$, and the corresponding standard deviations drawn from $U(1, 10)$. The ranges of the means and standard deviations are designed to provide sufficient amount of variation among exposures both within and between the areas. The average values for each area calculated using all the simulated exposures, $\widetilde{Z}_i = \sum_{j=1}^{N_{s_i}} Z_{ij}/N_{s_i}, i = 1, \ldots, 100$, are used to generate yearly mortality counts at each grid:

$$\begin{aligned} Y_i^{(1)} &\sim Poisson(\mu_i), \text{ for } i = 1 \ldots 100, \\ \log(\mu_i) &= \beta_0 + \beta_1 \widetilde{Z}_i^{(2)} \end{aligned} \tag{10.1}$$

where β_1 is the log of the true relative risk which is chosen to be RR$= 1.1$, reflecting the magnitude of the risks commonly observed in previous studies (Elliott et al., 2007) with the intercept term, β_0, is chosen as -2 in order to ensure that the simulated health data is within

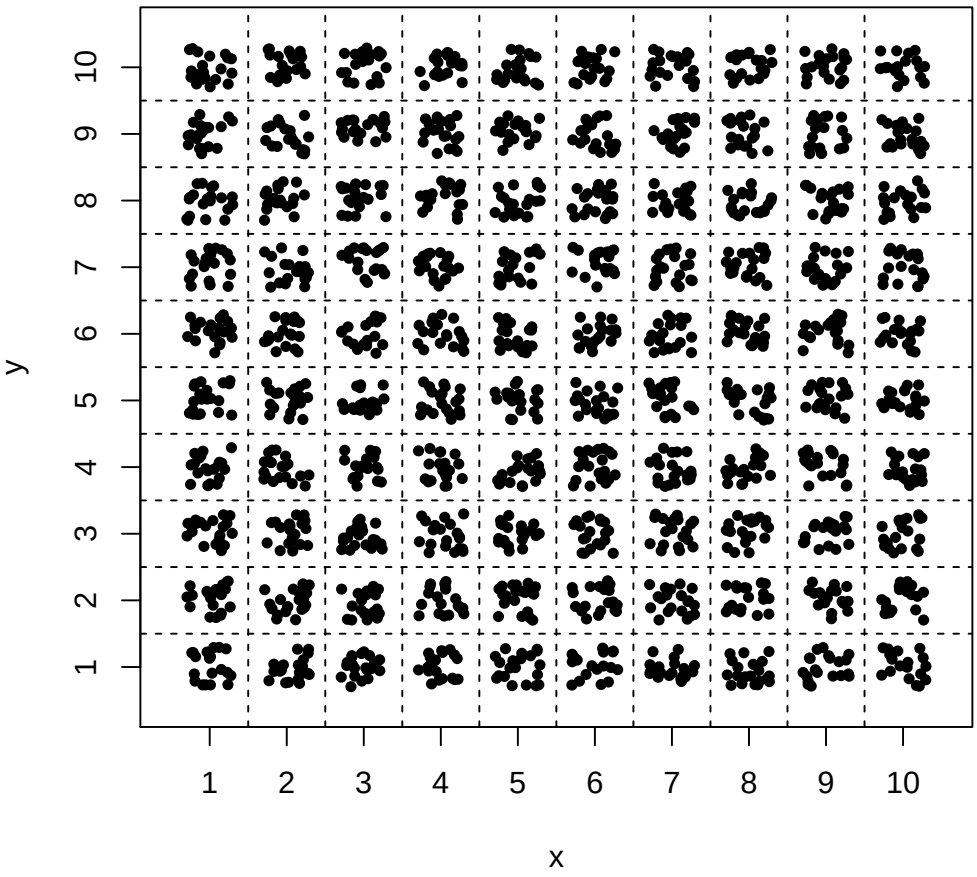

Figure 10-1: The 10×10 lattice used to generate data in the simulation study.

a suitable range. The key of this simulation is using a fixed single set of mortality data and using different sets of exposures induced under preferential sampling. Three sets of exposure values are considered in which the values calculated for each area are based on different sets of monitoring sites.

Set 1: No preferential sampling - using the average of all exposures in an area to represent the exposure in an area.

Set 2: Preferential sampling - using the average of the five highest values in each area to represent the exposure.

Set 3: Preferential sampling - using just the maximum value in each area to represent the exposure for the area.

For the simulation study, each of these are repeated 200 hundred times, generating 200 sets of exposures and corresponding health data. In each case, the the Poisson health model (10.1) is fitted generating 200 estimates of relative risk. The results from the simulation studies are shown in Figure 10-2 and 10-3 which show the estimated relative risks together with 95% confidence intervals for each of the 200 datasets using each of the three sets of exposures listed above. In the first of these the intercept term is fixed to be the true value (-2) used in generating the data whereas in the second both the intercept (β_0) and the estimate of risk (β_1) are estimated from the given data.

Results

The results of preferential sampling can be clearly seen in Figure 10-2 in which the true relative risk (1.1) is obtained when using all the available data in each area to represent the exposure. When the average of the five highest values is used, the estimated relative risks are decreased (the median value over the 200 datasets is 1.087) with a further decrease observed (median 1.080) observed when using the maximas. When the intercept term, β_0 is not fixed, as in Figure 10-3 again decreases in the relative risks are observed, but here the effects are driven by a similar mechanism to modelling with (classical) measurement error, this will result in attenuation of the risk parameter and an increase in the estimate when there is no exposure.

10.2 Statistical Modelling

10.2.1 Modelling the health effects of air pollution

As described in Chapter 9, Section 9.3, Poisson regression may be appropriate when the dependent variable , $Y_i^{(1)}$, is counts of disease arising from a defined population, N_i. In addition

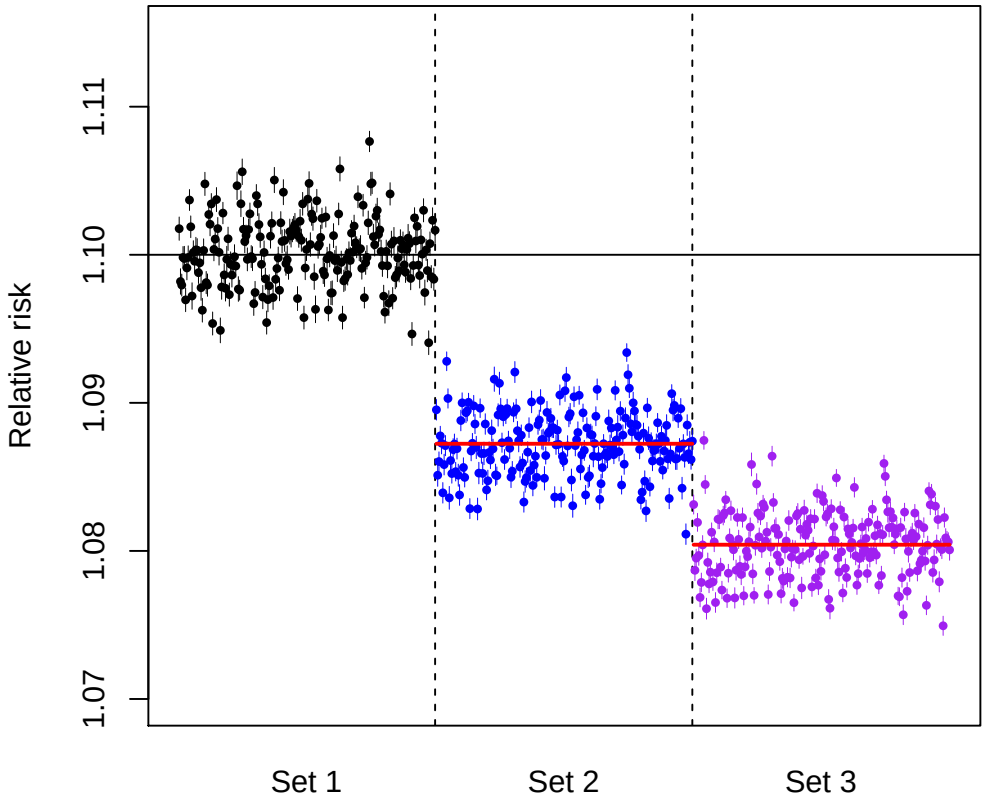

Figure 10-2: Results from applying health model to three sets of exposures representing different levels of preferential sampling: Set 1 using all available data; Set 2 using the highest 5 exposures in each area and Set 3 using the maximum value in each area. Dots represent the estimated relative risks and vertical lines the associated 95% confidence intervals. The horizontal line shows the true value of the relative risk used in the simulation, 1.10. In this example, the value of the intercept is fixed to be equal to the true value, -2.

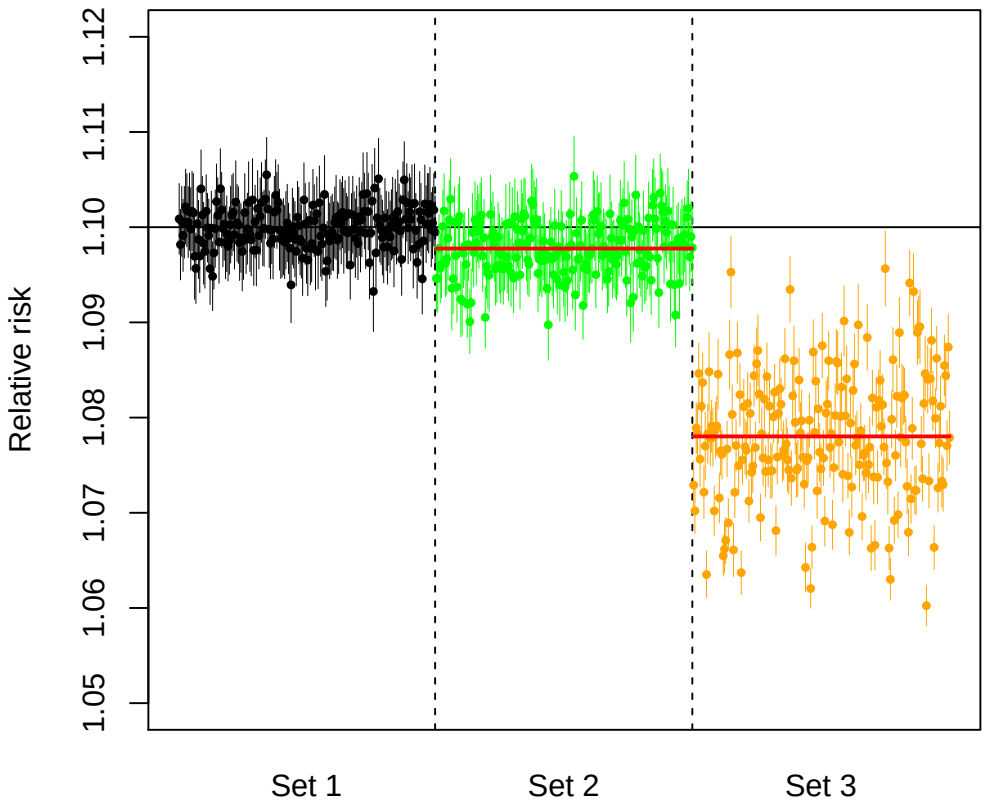

Figure 10-3: Results from applying health model to three sets of exposures representing different levels of preferential sampling: Set 1 using all available data; Set 2 using the highest 5 exposures in each area and Set 3 using the maximum value in each area. Dots represent the estimated relative risks and vertical lines the associated 95% confidence intervals. The horizontal line shows the true value of the relative risk used in the simulation, 1.10.

to just the population at risk, the number of observed cases will be determined by the age–sex profile of the population of interest and for this reason age–sex standarised *expected* numbers are often calculated in order to ensure that, as much as possible, any comparison and differences are due to exposures and not to underlying population characteristics (see Section 9.3 for further details).

Considering a generic area/time period of interest, i within those under study $i = 1, ..., N$, if $Y_i^{(1)}$ denotes the number of health outcomes in area i, it is assumed that $Y_i^{(1)} \sim P(E_i\mu_i)$, where the rate, μ_i is the expected number of counts adjusted by the risk for that particular area. The rate for area is modelled as a function of the exposure of interest for that area, $Z_i^{(2)}$, together with other area-level covaraites, $X_{pi}^{(1)}$,

$$\log \mu_i = \text{offset}(\log(E_i)) + \beta_0 + \beta_1 Z_i^{(2)} + \sum_{p=1}^{P} \beta_p X_{pi}^{(1)} \tag{10.2}$$

where $\text{offset}(\log(E_i))$ is a known multiplier, β_1 represents the effect of the exposure of interest, $Z_i^{(2)}$, and β_p is the effect of other area-level covariates, $X_{pi}^{(1)}, p = 1, ..., P$. The relative risk (RR) associated with a (unit) change in air pollution is RR$= \exp(\beta_1)$.

As described in Section 9.5, health counts are routinely available in aggregated form for defined administrative areas, A_i, whereas exposure data is measured at a number of fixed monitoring locations from N_s fixed site monitors located across the total area under study, $\mathcal{S}$. The set of pollution locations are collectively denoted by $S = s_1, \ldots, s_{N_s}$ (where $s_l = (a_l, b_l) \in \mathcal{R}^2$) with associated measurements denoted by $Y_s^{(2)}$. In order to obtain exposures at the same level of aggregation as the health data, exposure measurements need to be aggregated within health areas to produce a useable measurement in the health model 10.1;

$$Z_i = \int_{s \in A_i} N_i Z_{st}^{(2)} ds \tag{10.3}$$

where $Y_{st}^{(2)}$ are the ambient pollution concentrations measured at locations s at time t. N_i is the population density such that $\int_{s \in A_i} N_s ds = 1$. However the required information to perform the integral (10.3) will usually be unknown, and two general approaches have been adopted to obtain aggregated exposures. The simplest approach, which is used by the majority of studies, is to simply average the measurements in each of the health areas, i;

$$\widetilde{Z}_{it} = \overline{Y}_{it} = \frac{1}{N_{s_i}} \sum_{s \in A_i} Y_{st}^{(2)} \tag{10.4}$$

where N_{s_i} is the number of monitors in health area i. The use of this simple average, which ignores missing values, is due to its simplicity and ease of computation. However, there is no allowance for the fact that pollution concentrations can vary widely over space, i.e. within the health areas, due to the heterogeneous nature of pollution sources. For example, in urban areas particulate matter (PM) is predominantly produced from vehicle emissions and thus its concentration at any given location will depend highly on the local traffic density. Pollution fields are also continuous by nature and not subject to artificial boundaries imposed on them by arbitrary areas defined by health geographies.

10.2.2 Preferentially sampled exposures

We now consider the case where a subset of the monitoring locations are preferentially sampled. We assume that a subset of the monitoring locations, S_1, are not preferentially sampled and produce unbiased estimates of the exposures being experienced by the populations at risk. For the others, we consider the most probable case that the set $S_2 = S \setminus S_1$ are located where exposures are likely to be high. When site locations are allocated to health areas and the exposures aggregated, this will lead to the definition of two sets of health areas, H_1 in which only non-preferential sampled locations contribute exposure information, $Y^{(2)}_{S_1}$, and H_2 containing areas where the exposures $Y^{(2)}_{S_2}$, at least in part, come from preferentially sampled sites.

Considering areas H_1 based on the unbiased exposures, $Y^{(2)}_{S_1}$, it is assumed that aggregation of exposures, $Z^{(2)}_i$, in areas that contain these monitoring locations, are also unbiased. For simplicity we ignore the possible issues of measurement error and ecological bias will may arise when using exposures based on spatial predictions or when performing aggregation. For further details on the effect of spatial prediction see Szpiro et al. (2011), Chang et al. (2011), Lee and Shaddick (2010), then Haneuse and Wakefield (2007), Haneuse et al. (2008), Wakefield and Sebastien (2008) for approaches which directly address the issues of ecological bias.

Retained and non–retained sites

During the period of monitoring, at any point in time we want to determine two groups of sites that may as closely as possible represent non-preferentially (NPS) and preferentially sampled (PS) locations. Ideally this split would be involve detailed knowledge of the design of the network but in practice such information may be of limited availability. However it may be possible to create groupings based on changes of the network over time. Shaddick and Zidek (2014) analysed data for 1966-1996 from the network considered here, showed there were differences in levels of pollution and changes over time for four distinct groups: (i) sites which

were opertational throughout the entire period of study; (ii) sites that were dropped from the original network; (iii) sites that were added and (iv) sites that were added and subsequently dropped. Zidek et al. (2014) also investigated this, modelling future inclusion in the network as a function of previous measurements. In both cases, there was evidence that sites were being retained, and added, in locations where concentrations were high and sites were dropped when measurements were low. Using this information, two groups could be defined representing preferential sampling; those containing sites that were retained, S_1, (groups (ii) and (iv)) and those non-retained S_2, ((i) and (iii)). Although this is not going to be a perfect distinction between PS and NPS sites, it can be argued that any misclassification is likely to be conservative in that for example S_1 will contain some NPS sites, especially if the entire network tends towards being preferentially sampled, and so any observed differences will be smaller than those that may actually exist.

The definition of retained/non–retained sites and thus the inclusion in S_1 or S_2 is defined over a specified time period which, for area i, will be,

$$R_i = \begin{cases} 1 & \text{if } I(Y_{iT}^{(2)} > 0) = 1 \\ 0 & otherwise \end{cases}$$

where T represents the last year of the period of study, and the $R_i = 0$ indicates membership of S_1 and $R_i = 1$ of S_2. $I(Y_{iT}^{(2)} > 0)$ is an indicator term which is equal to one if there is a data point for the last year within the period of interest and zero otherwise. It is noted that the condition $Y_{iT}^{(2)} > 0$ is used to reflect that a value of zero is highly likely to be a missing data point rather than an actual reading of zero, which in this case would be beyond the possible accuracy of the monitoring equipment.

Having defined the two groups of sites, we now turn to the process of aggregation of point (site) level exposures to the health areas, which are now defined as H_1 (non-retained) and H_2 (retained) depending on the classification of the sites within them. In cases where the health area contains only sites in S_1 or S_2 it will be classified as H_1 or H_2 respectively, where there is a mixture $S_1 \& S_2 \rightarrow H_2$. Therefore if any site in a health area is retained, then there will be a continuous set of exposure data and thus the health area is classified as retained. This means the second group, H_2 will contain exposures which include information from sites in S_1 albeit combined with information from at least one site in S_1, which will result in a reduction in any difference between the two groups and evidence that if differences in risk between the two are observed then any underlying difference is likely to be greater in reality.

10.2.3 Predicting exposures

Differences between the relative risks, RR_1 and RR_2, estimated from using $Y_{S_1}^{(2)}$ and $Y_{S_2}^{(2)}$ would be indicative of an effect of preferential sampling. In addition, we may be able to predict the values for $Y_{S_2}^{(2)}$ as though they were in S_1 by using predictions from a spatial model for the exposures. This may be achieved using a hierarchical model with three components; (i) a spatial model for concentrations at non-preferentially sampled sites, named the 'group 1' model, (ii) prediction from this model at preferentially sampled locations and (iii) using the combination of the data from group 1 with the predictions for group 2 in a health analysis.

Stage one: group 1 model

The aim of this stage of the model is to estimate the spatial structure of the non-preferential sites together with the effects of any covariates. Let $Y_{st}^{(2)}$ represent the log transformation of the concentrations measured at group 1 sites, $s \in S_1$, at time t. We use the form of the exposure model seen in Equation 9.13 and apply it only the sites in the first (NPS) group. The modelling is performed on the log scale and so predictions will be transformed back to the original scale for the health modelling.

$$Y_{st}^{(2)} = (\beta_0^{(2)} + \beta_{0s}^{(2)}) + (\beta_x^{(2)} + \beta_{xs}^{(2)})t + (\beta_{x^2}^{(2)} + \beta_{x^2s}^{(2)})t^2 + \epsilon_{st} \tag{10.5}$$

where $s = 1, \ldots, N_S$ denotes the site and $t = 1, \ldots, N_T$ the year. The model includes both linear and quadratic effects, $\beta_x^{(2)}$ and $\beta_{x^2}^{(2)}$ of time reflecting the shapes of decline in the decline in levels of black smoke observed in the data. The ϵ_{st} is a random error term, which is assumed to be Normally distributed, $\epsilon_{st} \sim N(0, \sigma_\epsilon^2)$. Site specific random effects, $\beta_{xs}^{(2)}$ and $\beta_{x^2s}^{(2)}$ and $\beta_{0s}^{(2)}$, are assigned to the slopes of the linear, quadratic and intercept components respectively. These are contained to sum to zero, around fixed effects, $\beta_0^{(2)}$, $\beta_x^{(2)}$ and $\beta_{x^2}^{(2)}$ respectively. After allowing for the effects of time, there is likely to be spatial structure in the residuals and therefore the random effects are multivariate normally distributed, $\beta^{(2)} \sim MVN(0, \sigma_s\Sigma)$, with the structure of the covariance reflecting any spatial auto-correlation.

Stage two: prediction at group 2 locations

If the random error terms, ϵ_{st}, in (10.5) are uncorrelated, then a prediction at a new location, s' will comprise the combination of the overall predictions of the spatial terms at that new location. Here, the random effects for intercept and slopes will all be predicted at the new location. To explain this we consider a single set of random effects, $m = (m_1, \ldots, m_{N_s})$ with spatial structure. A prediction at location j will be of the form:

$$\hat{Y}^{(2)}_{s'} = \hat{\beta}_0 + \sum_{k=1}^{G} \hat{\beta}_k X^{(2)}_{ps'} + \hat{m}_{s'} \tag{10.6}$$

This can be viewed as two separate process; the first predicting covariate effects at group 2 locations and the second predicting the spatial effects. The spatial component is calculated using properties of the multivariate normal distribution. If $m = (m_1, .., m_{N_s})'$ are the observed values at the monitoring locations, then the conditional distribution of $m_{s'}|m$ at a new location, s', will be normally distributed with mean and variance given by

$$E[m_{s'}|m] = \sigma_m^{-2} \delta'_{s'} \Sigma_m^{-1} m, \tag{10.7}$$

and

$$Var(m_{s'}|m) = \sigma_m^2 - \delta_{s'} \sigma_m^{-2} \Sigma_m^{-1} \delta_{s'}, \tag{10.8}$$

respectively, where $\delta_{s'}$ is the vector of distances between the new location and the monitoring sites and $\delta_{s'} = f(d_{ss'}, \phi)$.

We now have two sets of exposures for the sites in S_2; the original data, $Y^{(2)}_{S_2}$ which is suspected to be subject to preferential sampling and now $\hat{Y}^{(2)}_{S_2}$, predicted from the model based on data from S_1. The interest is in whether the relative risks obtained using data from S_1 and S_2 differ and also in the results using $Y^{(2)}_{S_2}$ and $\hat{Y}^{(2)}_{S_2}$.

Stage three: combining data and predictions for use in health model

A set of data for all locations in S is now produced by collating the data from the sites in S_1, $Y^{(2)}_{S_1}$, with the predictions at the locations in S_2, $\hat{Y}^{(2)}_{S_2}$, with the average taken for each area (from Equation 10.4) now being

$$\widetilde{Z}_{it} = \sum_{i=1}^{N_i} \frac{(Y^{(2)}_{st} I_{st} + \hat{Y}^{(2)}_{st}(1 - I_{st}))}{N_i} \tag{10.9}$$

where I_{st} is an indicator variable which is one if site $s \in S_1$ and zero if $s \in S_2$ and N_i is the number of sites within health area i.

10.2.4 Inference

Treated on its own, the Poisson model is a standard GLM. There is a strong possibility of over-dispersion in the Poisson models (i.e., where the variance is greater than the mean) arising from the presence of unmeasured confounders. These may be operating at the individual level, e.g. smoking, or at the area level, e.g. residual socio-economic confounding. Over dispersion may also arise because of data anomalies, i.e. errors in the numerators and/or denominators, e.g. due to migration. Quasi-likelihood can be used to allow for extra-Poisson variability (McCullagh and Nelder, 1989). A Bayesian implementation of (10.2) could be random effects to be used to accommodate the over–dispersion.

As described in Section 9.5, here we fit the exposure and health models within a two–stage framework. Here the exposure models are implemented using Integrated Nested Laplace Approximation (INLA) with samples from posterior distributions used in the health models using multiple imputation (as described in Section 4.3).

10.3 Case study

We use data from the black smoke network described in Section 9.6. From a peak in 1971 when it comprised of over 1200 sites over time it has reduced in size as levels of black smoke have declined to 220 sites in 1996 until in 2006, when it ceased operation, there were only 65 sites. Over time, many sites have been moved or replaced in order to reflect changing patterns and levels of pollution, and to reduce redundancy in the network. Therefore there is the possibility of selection bias if the monitoring sites are kept in polluted areas. Since 2006, black smoke has continued to be monitored at 20 locations as part of the Black Carbon network in order to provide a continuous source of historical information on levels of BS.

The BS network provides a unique record of pollution over a long period and has been used as the basis of epidemiological studies of the long term effects of air pollution on health (Elliott et al., 2007) and as the basis of studies estimating exposures of air pollution across the UK over several decades (Gulliver et al., 2011). Specifically, it has been used to estimate exposures over extended periods for health analyses (Morris et al., 2007). However, in these cases, no information on the choice of sampling locations or the effects of changes in the network over time have been considered. There is therefore a real need to understand the possible effects that the choice of locations of monitoring sites included in the network might have on the resulting estimates of exposures and further on the estimates of health risk which will arise from them.

Here we consider a unique period in history during which concentrations fell dramatically

from levels which would be unrecognisable in the UK today, reflecting changes in the large scale use of fossil fuels. As reported in Shaddick and Zidek (2014), annual means fell from 237 μgm^{-3} in 1962 to 99 in 1966, 32 in 1976 and 5 μgm^{-3} in 2006. In order to investigate the possible effects that preferential sampling may have on the estimation of health risks, as in Shaddick and Zidek (2014), we choose the time frame 1966-1996 for the analysis that follows, during which time dramatic changes in the network were observed which are likely to a large extent to have been a result of preferential sampling (see Shaddick and Zidek (2014)). The health data consist of mortality counts within 361 small areas (wards) for respiratory diseases (ICD9, 460-519) in the over 65s during 1981-1984.

10.3.1 Exposure model

We now describe the model used to predict the concentrations of the sites in S_2 as though they followed the same patterns over space and time as those in S_1 (see Section 10.2.3). Let $Y_{st}^{'(2)}$ be the concentration of black smoke measured at location, s, at time, t. In order to non–dimensionalize our measurements, we divide them by 78 (units), roughly the level of black smoke concentrations at the start of the period of study (see Shaddick and Zidek (2014) for further details). The unit less ratio, now represents the number of baseline units of decline in that particulate concentration since that time. Here then $Y_{st}^{(2)} = log(Y_{st}^{'(2)}/78)$.

Following the approach developed in Shaddick and Zidek (2014) and used in Chapter 9, the response is modelled as follows;

$$Y_{st}^{(2)} = (\beta_0^{(2)} + \beta_{0s}^{(2)}) + (\beta_x^{(2)} + \beta_{xs}^{(2)})t + (\beta_{x^2}^{(2)} + \beta_{x^2s}^{(2)})t^2 + \beta_U U_s + \beta_g G_s + \epsilon_{st} \quad (10.10)$$

where $s = 1, \cdots, N_s$ denotes the site and $t = 1, \cdots, T$ the year. The model includes both linear, β_x, and quadratic, β_{x2}, effects of time reflecting the shapes of decline in levels of BS. Site specific random effects, β_{0s}, β_{xs} and β_{x2s}, are assigned to the intercepts and slopes of the linear and quadratic components respectively. These are constrained to sum to zero, around fixed effects, β_0, β_x and β_{2x} respectively. The effects of a site being located in a rural area are represented by β_u with U_s being an indicator variable reflecting whether a location was rural or not. The type of a site is specified by the indicator G_s. As detailed in Section 10.2.2 we define two groups which are intended to as much as possible to reflect differences in non-preferentially located (Group 1; G_1) or preferentially located (Group 2; G_2) which are meant to represent S_1 and S_2 respectively. Here, these are defined as follows:

G1: Non-retained - sites that were dropped from the network, including those that were added and subsequently dropped.

G2: Retained - sites that were consistently operational throughout the period of study or were added to the network during that time.

The ϵ_{st} is a random error term, which is assumed to be Normally distributed, $\epsilon_{st} \sim N(0, \sigma_\epsilon^2)$. In addition, interactions between the slope terms (linear and quadratic) and the group indicator are included, which allows for both a shift in overall levels between the two groups (retained and non-retained) and different rates of decline over time. The random effects terms are multivariate normally distributed, $\beta_{\mathbf{s}} \sim MVN(0, \sigma_s \Sigma)$, with the structure of the covariance reflecting any spatial auto-correlation as described in Chapter 2, Section 2.1.

10.3.2 Health model

Here we define the two groups of sites, which are intended to represent the NPS and PS locations by future changes in the network (after the period of health data that is being considered). The health areas (wards) were split into two groups, denoted H_1 and H_2, according to whether they contained non-retained or retained sites as described in Section 10.2.2.

We consider two approaches to the analysis which correspond to the use of different sets of exposures in the health model.

S 1 Using all available data, missing values of exposures for a particular year in an area results in the average (over time) for that area being based on a smaller number of data points.

S 2 Using predictions from an exposure model based on S_1 for locations in S_2. This also fills in the missing values in group 1 with predictions from the model.

For each of these scenarios, we fit the health model to the health and exposure data for the two different groups, H_1 and H_2, and also produce a combined result over both groups. We use the average exposure over the previous four years (1977-1980), representing an extended period of exposure but one which is short enough to reduce the possible issues of population migration (Elliott et al., 2007). Interest lies in whether there are differences in the resulting estimates of relative risks which would suggest that preferentially sampling of exposures can have an effect in health studies.

10.3.3 Results

The health groups, H_1, H_2, are defined by what happens after the time of the health analysis. Here, we use health data from 1981-1984 and define H_1 and H_2 based on what happens within the network from 1981 onwards, with strong evidence that the choice of which sites remained

was linked to levels of pollution at those locations (Shaddick and Zidek, 2014). This resulted in 300 wards in H_1 and 61 wards in H_2.

The locations of the monitoring sites in the two groups can be seen in Figure 10-4. Figure 10-5 shows the mean concentrations over all wards by year for the set of wards in H_1 (red) and H_2 (blue). There is a marked decline in the level of BS over the this period and a clear difference between the ward level exposures in the two groups.

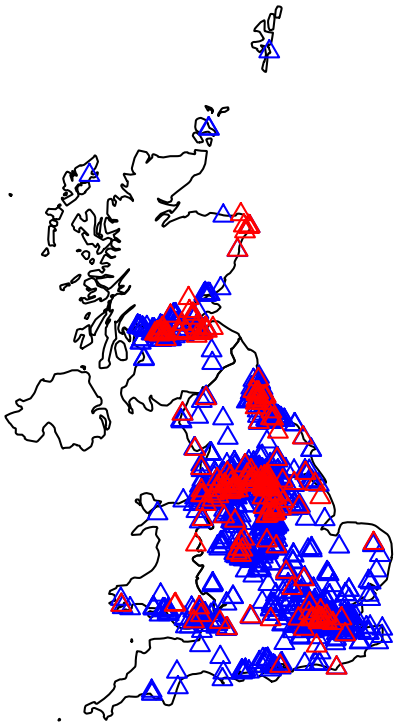

Figure 10-4: Locations of the sites in non-retained (blue triangles) and retained (red triangles) groups.

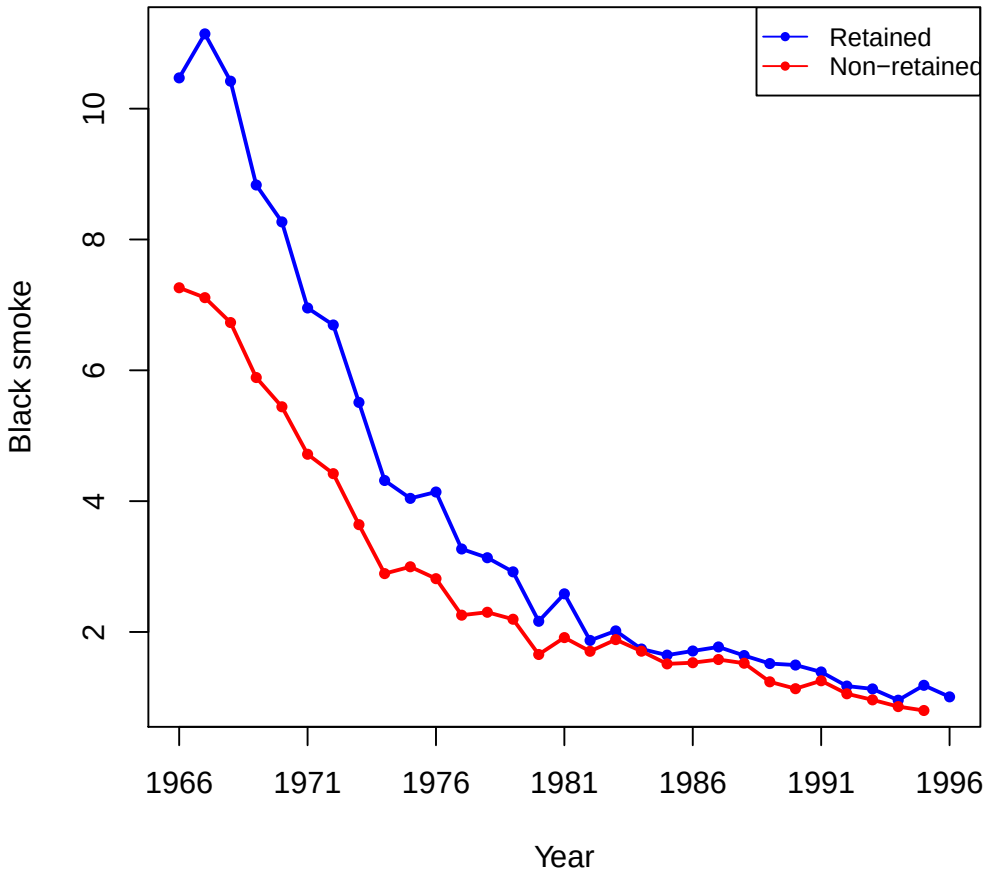

Figure 10-5: Mean concentrations of black smoke (over all areas) by year for the set of areas (wards) containing sites that were not retained, H_1 (red), and those for which the sites were retained, H_2 (blue) for the second set of analysis (using health data from 1981-1984).

For the second set of analyses, we use predictions from an exposure model based on S_1 for locations in S_2. The idea is that the measurements for sites which are considered to be preferentially sampling are replaced by predictions from the exposure model under the assumption that they are non–preferentially sampled, and thus as though they were in group S_1. A comparison of the predictions under this scenario and the actual measurements for the sites in S_2 by year can be seen in Figure 10-6. As might be expected, the predictions as though the sites were non–preferentially sampled are on the whole less than the actual measurements suggesting that the 'correction' is in the right direction. The corresponding graph showing the average over the 16 years can be seen in Figure 10-7 where again the 'corrected' measurements are smaller than the actual ones which is good because preferentially sampled exposures would be expected to be higher than they should be and now we will be using lower (adjusted) exposures in the health model.

Table 10.1: Relative risks (RR) of respiratory mortality, with 95% confidence intervals (CI) for increase of 10 ppb of black smoke over the previous 4 years (1977-1980) analysed in two scenarios: S1 – observed exposure values only and S2 – with measurements for group 2 estimated using predictions from a spatial model based on data from group 1. Results are for two groups separately and combined and for S2 come from multiple imputation using 100 datasets using samples from the posterior distribution of a spatio–temporal exposure model (see text for details). Confidence intervals are given based on Poisson likelihood and quasi-likelihood to reflect possible extra-Poisson variability.

S1: Observed exposures

	Group 1		**Group 2**		**Overall**	
	RR	95% CI	RR	95% CI	RR	95% CI
Poisson	1.116	1.096–1.136	1.068	1.040–1.098	1.075	1.061–1.089
Quasi	1.116	1.039–1.120	1.068	0.992–1.151	1.075	1.023–1.128

S2: Predictions from group 1

	Group 1		**Group 2**		**Overall**	
	RR	95% CI	RR	95% CI	RR	95% CI
Poisson	1.116	1.097–1.137	1.089	1.053–1.125	1.088	1.072–1.102
Quasi	1.116	1.096–1.137	1.089	1.000–1.180	1.088	1.032–1.146

The results of the health analyses can be seen in Table 10.1. In the first set of results, the observed measurements of concentrations are used as would be used in a traditional epidemiological analysis. The overall relative risk (RR) is 1.075 which is significant (95% CI; 1.061–1.089). This remains significant when using a quasi–likelihood model although the confidence interval gets wider. Within this analysis, splitting the wards into the two groups, H_1 and H_2 shows that the RR in the first (non–preferentially sampled group) is greater than that the preferentially sampled group (1.116 vs. 1.068). In the second set of analyses we use predictions from

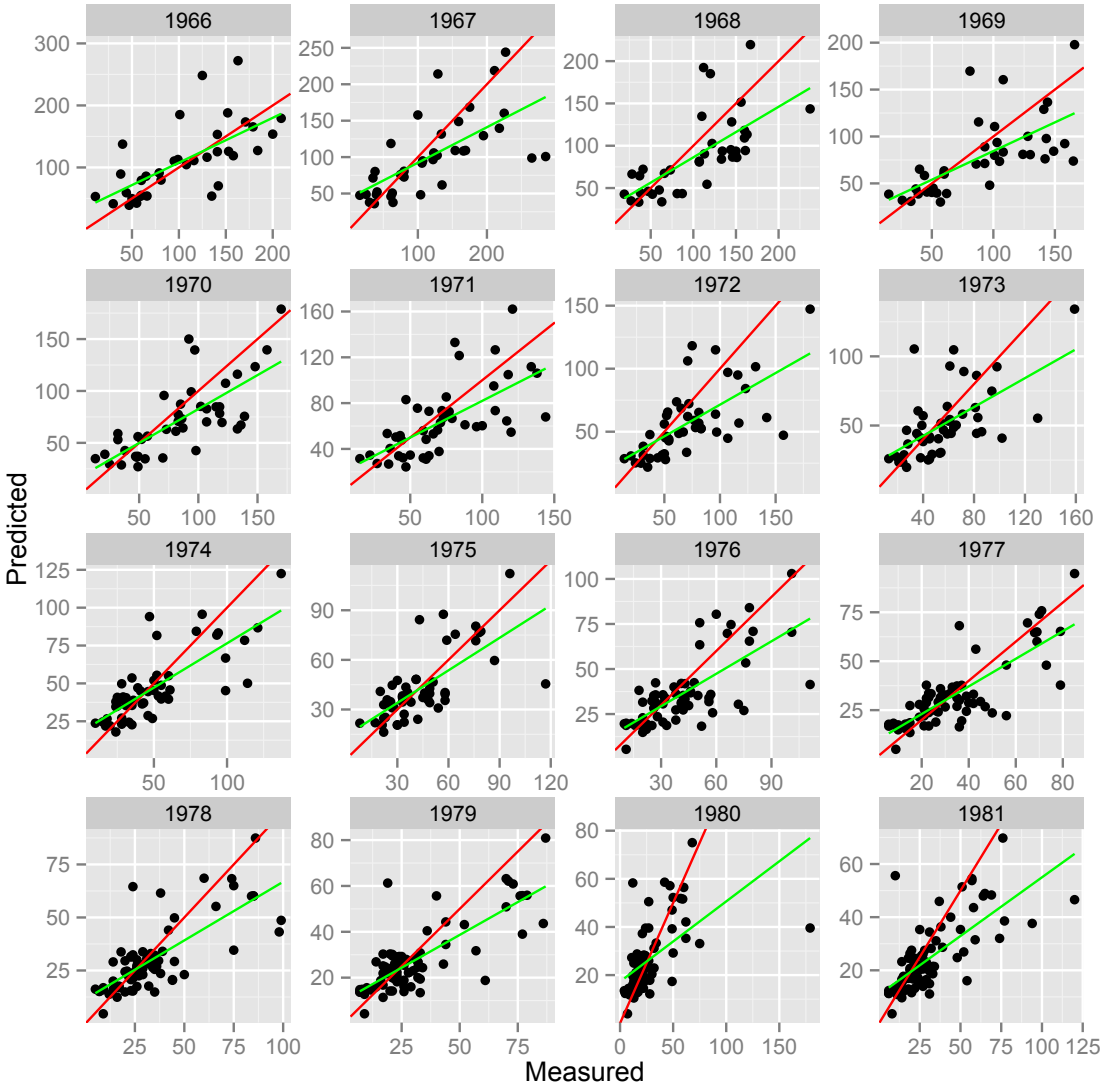

Figure 10-6: Predicted and measured values of concentrations in sites that are retained (see text for details) for 1966 - 1981. In each panel, the red line has zero intercept and a slope of one whilst the green line is the line of best for through the data.

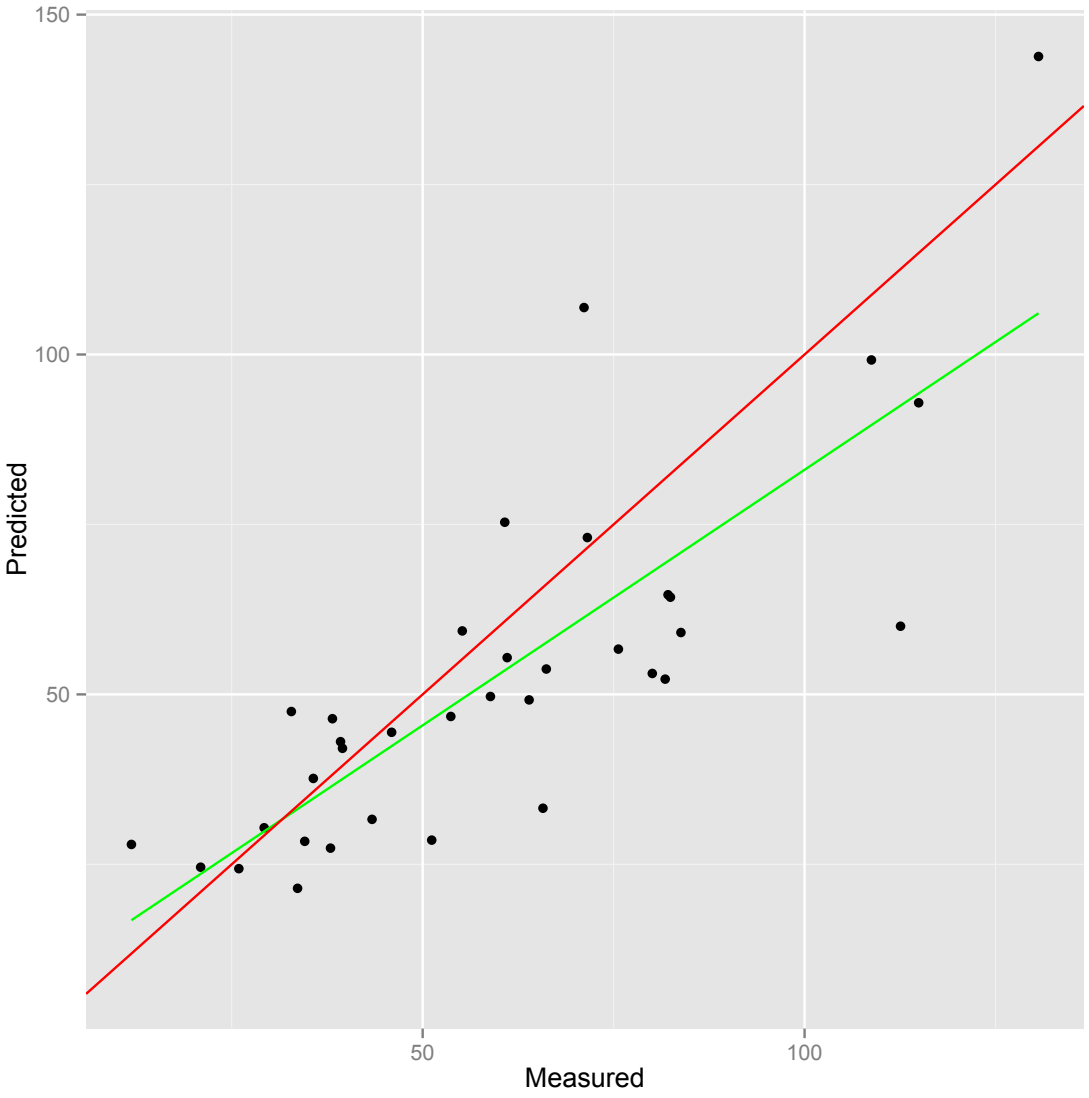

Figure 10-7: Predicted and measured values of concentrations in sites that are retained (see text for details) averaged over the sixteen years from 1966 to 1981. In each panel, the red line has zero intercept and a slope of one whilst the green line is the line of best for through the data.

a model in place of the data for the second group. The predictions are made under the assumption that the sites in the second group are actually in the first, i.e. they are non–preferentially sampled. As seen in Figures 10-7 and 10-6 the exposure values used here will be smaller than in the first analysis and so we would this 'correction' to result in higher RRs which should be closer to those observed from the non–preferentially sampled group. This is indeed the case and the RR is now 1.089 which is greater than the 1.068 previously seen and closer to the value seen in the non–preferentially sampled group. The overall result, combining data from both groups is increased from 1.075 (when real data is used for both groups) to 1.088 (when the predictions are used in place of data for the second group). The final result, including the 'correction' is therefore 1.088 (95% CI; 1.032–1.146 using quasi–likelihood) which indicates a significant increase in risk associated with higher long–term exposures to air pollution

10.4 Conclusion

In this chapter we investigated the effects that preferential sampling can have on the estimates of relative risks. The phenomenon preferential sampling occurs when monitoring sites are located as a consequence of the network design with certain purposes, e.g. only collecting high values. The air pollution data collected from such networks are unlikely to be a adequate representation of the pollution level in the whole study region. From a statistical point of view such data may not be suitable to use as proxies for the exposures that might be experienced by populations, and inherently may not provide accurate estimates of underlying exposures. Therefore, preferential sampling has the potential to introduce bias into estimates of risk associated with exposure to air pollution and subsequent health impact analyses.

In the the simulation studies, we show the preferentially sampled exposures are associated with different estimated relative risks compared to the ones related to overall exposures. Ideally, the health analysis can be conducted using the exposure measurements from a subset of the monitoring sites which can be assumed to be non–preferentially sampled, but the nature of monitoring networks means that there may be very few of these type of sites and so the sample size for the health analysis may well be too small. We proposed an approach to adjust the preferential sampling effects on relative risk using the exposures from a set of non-preferential sites to predict what might have been measured at preferentially sampled locations. This would result in using lower measurements of exposures in the health model for these sites which should result in an overall estimate of risk which is closer to that which would (theoretically) have been observed if the monitoring network was a fair representation of what might have been experienced by the population.

For the case study, we use future changes in the monitoring network to define non–preferential and preferential groups of site locations. The future information used to make the distinction between the two groups of sites was available at the time of analysis, but this may not be the case in other studies. It may be possible to predict suitable membership by predicting future changes in the network, i.e. whether sites are likely to be withdrawn or retained, using logistic regression. Furthermore, in this study and the case study in Chapter 9 we use a relatively simple random effect exposure model with quadratic form to represent temporal trend due to the change of the average black smoke concentration during 30 years. It may be worth to fit a more sophisticated model to represent temporal component of exposures in INLA, such as a autoregressive process.

Chapter 11

Discussion

Air pollution is an important determinant of health. There is convincing, and growing, evidence linking the risk of disease and premature death with exposure to fine particulate matter ($PM_{2.5}$) and ozone (O_3). The public health burden of present exposure is substantial. Recently published Global Burden of Disease assessments indicated that ca. 7 million premature deaths per year and 3.1% of the global disease burden could be attributed to ambient particulate matter pollution, placing it among the top health risk factors globally. It is therefore extremely important that the risks to health are estimated accurately and for that to happen there need to be accurate levels of the exposures that might be experienced by populations at risk. As air pollution is experienced by everybody, even small increases in risk can be associated with very large numbers of people being ill.

11.1 Health modelling

The standard approach to modelling health outcomes is to use a Poisson log-linear model relating measures of air pollution to counts of mortality or morbidity. To allow for extra-Poisson variability, where the Poisson assumption that the variance is equal to the mean is not correct, quasi–likelihood methods are often used. This results in larger estimates of the variance associated with estimated risks and thus confidence intervals are wider. This means that it becomes more likely that increases in risk will actually be non–significant when the modelling assumptions are more reasonable. However, even when using quasi–likelihood there are assumptions that are commonly made that may mean that the estimated risks from studies are not entirely accurate. The most common way of using exposure measurements is simply to average them over an area for a specified period of time. In Chapter 8 this is seen in the context of short–term effects of air pollution in which measurements from a number of monitoring sites within an area (the city of London in this case–study). Unfortunately just taking an average and using it

in the health model ignores the fact that it is really a summary of a number of data points. It therefore ignores the uncertainty that is associated with it and really this should be acknowledge in the health model. This uncertainty should be reflected in the confidence intervals around the estimates of risk. Ignoring it in the modelling process may mean that confidence intervals are too narrow and so results may be seen as significant when really they should not be.

In Chapter 4 we investigated how well the standard model, using the average of measurements over monitoring sites for each day, did in the presence of underlying variability. If there is no underlying variability then there will be no problem, all the measurements in a day will be exactly the same and so the mean will be a true representation of pollution levels. However this is very unlikely in reality where there will be changes in air pollution over short distances, especially in urban areas where things such as buildings and roads can have a great effect on the levels recorded. In such cases, a model should ideally allow for such variability but in the standard model this is ignored. Through extensive simulation studies, we assessed the bias in relative risks that will arise if the model is too simplistic. This pointed towards the need for more complex models which allow the underlying variability to be modelled.

11.2 Exposure modelling

In Chapter 6 we propose the use of a Bayesian model which allows the exposures to vary over space. A spatio–temporal model is presented which we fit using MCMC. Using simulation studies, we showed that such a model was able to estimate the true risks to health in a much more accurate way than the standard model performed in Chapter 4. As well as a spatio-temporal model, we also looked at a simpler version where the underlying variability did not have any spatial structure, which we called a measurement error model, although a more accurate name might be a non-spatially structured model. Both of these were able to accurately estimate the true relative risks in the simulation study although the confidence intervals were slightly narrower when using the spatial model. However, we think there are also advantages to using the simpler, measurement error, model in that it is less computationally demanding. Also there may be issues with convergence of the parameters of the spatial model (Finley et al., 2007), especially if datasets are relatively small, in which case there might not be enough information to estimate them accurately.

Issues with the parameter controlling the relationship between distance and correlation were seen in both the simulations and the case study presented in Chapter 8. Notably, the posterior distribution seemed to be highly reliant on the choice of bounds in the prior. In the example shown in Chapter 8 this suggested that the prior for the spatial correlation parameter

should have been wider however when performing sensitivity to prior choice it was found that changing the upper bound resulted in the same effect. This suggested that the value of the parameter might be even greater than estimated, with large values of this parameter indicating very small correlation over even small distances. If it is the case that there is not really any spatial structure in the data then fitting a complex spatial model would be unnecessary and in fact might even introduce bias. Of course it may be that the spatial structure does not really fit into the class of models being considered and so the model will be misspecified which also may lead to bias which may well be fed though to create bias in the estimates of health effects.

However, a spatial model can be assumed to be a good fit if it has a number of advantages over a model which does not impose structure on the underlying variability. Firstly, it is important to have as much of the variability in the data explained by the model. Ideally this would be using covariate information but after that understanding the nature of the residual variability is better than just saying it is random and unexplained. Spatial variation in this case is saying that there are unmeasured confounders which have some spatial structure and that our model is acting as a proxy for these. Also, modelling the spatial structure allows us to perform predictions of air pollution for locations where data is not available. As shown in Chapters 9 and 10, this can be an extremely useful technique both for filling in missing data but also in reducing the bias that may arise when using the available data in its raw form. For example, in Chapter 9 we showed that where there are changes in the levels of pollution over time, misleading results can come from just ignoring missing values. If data are missing from a period when air pollution was high then summary measures may be too low, being based on data from when pollution was lower. If data are missing from times where pollution was low then summary measures that ignore missing data will be too high. Both cases have the potential to then affect the estimated risks to health. In Chapter 9 we showed how an exposure model can be used to 'fill in' missing values to reduce this type of bias. An additional advantage is that it also means the sample size is bigger. However, in using predictions from an exposure model as the inputs to the health model the fact that they are themselves modelled and are thus subject to uncertainty should be acknowledged in calculating confidence intervals.

11.3 Linking exposure and health models

In a fully Bayesian framework estimation of health and exposure models, including prediction at locations where data is not available, is performed simultaneously. The uncertainty in estimating the coefficients of the exposure model is therefore acknowledged and 'fed through' the model to the predictions and further to the estimation of the coefficients in the health model. However, there may be conceptual reasons why 'feedback' from the health model to the ex-

posure model is not desired. Here, it is the exposures that might be thought of causing health effects but the health effects are not thought to affect the exposures in the same way. It is noted that although an epidemiological regression model cannot itself prove causality, that can only really be provided by randomised experiments, it can indicate the change in the response variable that might be associated with changes in exposure, either by prediction or estimation, which is a very useful tool in developing insight and understanding into possible causal relationships.

There may also be computational considerations associated with jointly fitting the health and exposure models, especially if the latter uses large amounts of data over space and time. When the exposure model is complicated or when one is interested in running multiple candidate epidemiological models with different sets of covariates either for a single outcome or multiple outcomes, a single model is not going to provide an efficient method of investigation. A two-stage approach has the advantage that one does not have to rerun the exposure model when running multiple health effect analyses. Two stage approaches separate the exposure and health components, whilst still allowing uncertainty from the exposure modelling to be incorporated into the health model (Chang et al., 2011; Peng and Bell, 2010; Lee and Shaddick, 2010).

In theory, it would be relatively straightforward to fit the models considered in this thesis using MCMC and in this would provide a natural way of allowing the uncertainty associated with using predictions from the exposure model to be fed through to the estimates of the health risks. However, in practice the computational requirements may prove to be prohibitive, both because of the requirement to manipulate large matrixes within each simulation of the MCMC and also in convergence of parameters in complex models.

The models in Chapters 9 and 10 here were fitted using INLA with the SPDE approach to allow point referenced spatial components to be incorporated. In terms of prediction at a very high number of locations, techniques such as INLA, which perform 'approximate' Bayesian inference and thus do not require full MCMC sampling provide an extremely appealing approach, as shown in Lindgren et al. (2011). Overall, the implementation of the INLA and SPDE approaches in this thesis demonstrate how the methods can provide a remarkably fast computational algorithm for application over large domains when standard computational methods might fail.

11.4 Preferential sampling

Formulating guidelines with specific reference to health requires accurate information on the state of air pollution at different periods of time and over different areas which will be obtained from monitoring networks. However the information that is available to support air pollution policy and management is far from sufficient and three specific problems conspire to limit its utility: (i) monitoring is expensive and so monitoring networks are typically sparse; (ii) concentrations may vary greatly over small distances, especially in urban areas; and (iii) networks are often designed to monitor compliance with standards and may not give a true representation of levels over an area.

Commonly monitoring sites may be located where measurements might be expected to be high, a phenomenon referred to as *preferential sampling* and will be a consequence of the *network design*. The network may be designed to check adherence to standards but this may cause difficulties in epidemiological research. Measurements from such networks are unlikely to be representative of the pollution level in a wider area and, in their raw form, may not be suitable for use as proxies for the exposures that might be experienced by populations.

From a statistical point of view the data arising from preferentially-designed networks may not accurately characterise the spatio-temporal fields they intend to monitor and inherently will not provide accurate estimates of exposures. This has the potential to introduce bias into estimates of risk associated with exposure to air pollution and subsequent health impact analyses. Recent research has begun to address the issue of preferential sampling in environmental networks but there is little, if any, work in trying to assess or correct for the subsequent effects that it may have on health studies.

In Chapter 10 we investigated the effects that preferential sampling can have on the estimates of relative risks. Using simulation studies we showed that higher exposures were associated with lower relative risks and that this could well be an artefact that would be due to preferential sampling. Using the information from a monitoring network that is subject to preferential sampling, if there are a subset of the monitoring sites which can be assumed to be non–preferentially sampled then analysis might be restricted to just these. However in practice, the nature of monitoring networks means that there may be very few of these type of sites and so the sample size for the health analysis may well be too small. We have proposed an approach to reduce the effects on relative risks than involves using the information from a set of non-preferential sites to infer what might have been measured at preferentially sampled locations if they were actually not preferentially sampled. This would result in using lower measurements

of exposures in the health model for these sites which should result in an overall estimate of risk which is closer to that which would (theoretically) have been observed if the monitoring network was a fair representation of what might have been experienced by the population.

In the case study considered in Chapter 10, the two groups representing non–preferential and preferential site locations which generated the exposure data were defined based on a period of time after the epidemiological study. The aim was to use future changes in the monitoring network to indicate which monitoring sites should be in each group. In this case, the information required to make the distinction between the two groups of sites was known at the time of analysis which may not be the case in other studies. It may be possible to predict suitable membership by predicting future changes in the network, i.e. whether sites are likely to be withdrawn or retained, using logistic regression. Such an approach is proposed in Zidek et al. (2014) in the case of predicting an expanding network, although it could be adapted to this application, using spatial prediction of future events with multiple imputation to obtain measures of uncertainty.

11.5 Summary

In summary, the models developed and the conclusions drawn in this thesis should lead to more accurate estimates of the effects of air pollution on health. It has been shown that models that consider changes in exposures over time and space allow risks to be more accurately estimated. Such models also allow estimates of air pollution to be made during periods and in locations where there is missing data, either by design (where a monitor is not located or in operation) or due to shorter periods where measurements are not available. Such predictions can be used in health models to reduce the possible effects of bias due to missing data and as a potential methods for reducing the effects of preferential sampling. Furthermore these models should also lead to a greater understanding of the underlying processes that generate the pollution data, as well as the effects of commonly made assumptions can have on the statistical modelling process in general. There are however problems associated with using such complex models, the quantity of data required to produce reliable results is likely to be large, possibly more than is easily available in practice. Larger amounts of monitoring data are becoming increasingly available which will allow more accurate estimation but will come with an increased computational cost. In some cases the computation required may be prohibitively large, meaning that reliable estimates may still difficult to obtain. This may especially be the case with MCMC and therefore there is a need for possible alternatives producing approximate Bayesian inference such as INLA. Together with the general increase in computing power and data availability, these would enable increasingly realistic models to be considered. This will

lead to more accurate predictions of exposures in both time and space which will ultimately lead to more accurate estimates of the effects of air pollution on human health.

Bibliography

Bandeen-Roche, K., C. Hall, W. Stewart, and S. Zeger (1999). Modelling disease progression in terms of exposure history. *Statistics in Medicine 18*, 2899–2916.

Berger, J. O., V. De Oliveira, and B. Sansó (2001). Objective Bayesian analysis of spatially correlated data. *Journal of the American Statistical Association 96*(456), 1361–1374.

Berry, G., J. Gilson, S. Holmes, H. Lewinshon, and S. Roach (1979). Asbestosis: A study of dose-response relationship in an asbestos textile factory. *British Journal of Industrial Medicine 36*, 98–112.

Bornn, L., G. Shaddick, and J. V. Zidek (2012). Modeling nonstationary processes through dimension expansion. *Journal of the American Statistical Association 107*(497), 281–289.

Breslow, N. and N. Day (1980). *Statistical Methods in Cancer Research, Volume 2- The Analysis of Cohort Studies*. Scientific Publications No. 82. Lyon: International Agency for Research on Cancer.

Breslow, N., J. Lubin, P. Marek, and B. Langholz (1983). Multiplicative models and cohort analysis. *Journal of the American Statistical Association 78*(381), 1–12.

Breslow, N. E. and N. E. Day (1987). *Statistical methods in cancer research*, Volume 2. International Agency for Research on Cancer. Lyon.

Brezger, A. and S. Lang (2006). Generalized structured additive regression based on Bayesian P-splines. *Computational Statistics & Data Analysis 50*(4), 967–991.

Burnett, R. T., R. E. Dales, M. E. Raizenne, D. Krewski, P. W. Summers, G. R. Roberts, M. Raadyoung, T. Dann, and J. Brook (1994). Effects of low ambient levels of ozone and sulfates on the frequency of respiratory admissions to Ontario hospitals. *Environmental Research 65*(2), 172–194.

Cameletti, M., R. Ignaccolo, and S. Bande (2011). Comparing spatio-temporal models for particulate matter in Piemonte. *Environmetrics 22*(8), 985–996.

Carlin, B. P., H. Xia, O. Devine, P. Tolbert, and J. Mulholland (1999). Spatio-temporal hierarchical models for analyzing Atlanta pediatric asthma ER visit rates. pp. 303–320.

Carroll, R. J., D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu (2010). *Measurement error in nonlinear models: a modern perspective*. CRC press.

Carstairs, V. and R. Morris (1989). Deprivation: explaining differences in mortality between Scotland and England. *British Medical Journal 299*, 886–889.

Chang, H., R. Peng, and F. Dominici (2011). Estimating the acute health effects of coarse particulate matter accounting for exposure measurement error. *Biostatistics 12*(4), 637–652.

Chatfield, C. (2013). *The analysis of time series: an introduction*. CRC press.

Clifton, M. (1964). Air pollution. *Journal of the Royal Society of Medicine 57*(7), 615–618.

Daniels, M., F. Dominici, S. Zeger, and J. Samet (2004). The national morbidity, mortality, and air pollution study. PM10 concentration-response curves and thresholds for the 20 largest US cities. *Research Report (Health Effects Institute)* (94 Pt 3), 1–21.

Dewanji, A., M. Goddard, D. Krewski, and S. Moolgavkar (1999). Two stage model for carcinogenesis: number and size distributions of premalignant clones in longitudinal studies. *Mathematical Bioscience 155*, 1–12.

Diggle, P., R. Menezes, and T. Su (2010). Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 59*(2), 191–232.

Diggle, P. J., J. Tawn, and R. Moyeed (1998). Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 47*(3), 299–350.

Dockery, D. W., C. A. Pope, X. Xu, J. D. Spengler, J. H. Ware, M. E. Fay, B. G. Ferris Jr, and F. E. Speizer (1993). An association between air pollution and mortality in six US cities. *New England Journal of Medicine 329*(24), 1753–1759.

Dolk, H., G. Shaddick, P. Walls, C. Grundy, B. Thakrar, I. Kleinschmidt, and P. Elliott (1997). Cancer incidence near radio and television transmitters in Great Britain I. Sutton Coldfield transmitter. *American Journal of Epidemiology 145*, 1–9.

Dominici, F., J. M. Samet, and S. L. Zeger (2000). Combining evidence on air pollution and daily mortality from the 20 largest US cities: a hierarchical modelling strategy. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 163*(3), 263–302.

Duddek, C., N. Le, J. Zidek, and R. Burnett (1995). Multivariate imputation in cross-sectional analysis of health effects associated with air pollution. *Environmental and Ecological Statistics 2*(3), 191–212.

Eidsvik, J., B. A. Shaby, B. J. Reich, M. Wheeler, and J. Niemi (2013). Estimation and prediction in spatial models with block composite likelihoods. *Journal of Computational and Graphical Statistics 23*(2), 295–315.

Elliott, P., G. Shaddick, M. Douglass, K. de Hoogh, D. J. Briggs, and M. B. Toledano (2013). Adult cancers near high-voltage overhead power lines. *Epidemiology 24*(2), 184–190.

Elliott, P., G. Shaddick, I. Kleinschmidt, D. Jolley, P. Walls, J. Beresford, and C. Grundy (1996). Cancer incidence near municipal solid waste incinerators in Great Britain. *British Journal of Cancer 73*, 702–710.

Elliott, P., G. Shaddick, J. C. Wakefield, C. de Hoogh, and D. J. Briggs (2007). Long-term associations of outdoor air pollution with mortality in Great Britain. *Thorax 62*(12), 1088–1094.

Finazzi, F., E. M. Scott, and A. Fassò (2013). A model-based framework for air quality indices and population risk evaluation, with an application to the analysis of Scottish air quality data. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 62*(2), 287–308.

Finley, A., S. Banerjee, and B. Carlin (2007). An R package for univariate and multivariate hierarchical point-referenced spatial models. *Journal of Statistical Software 19*(4), 1–24.

Fuentes, M., H. Song, S. Ghosh, D. Holland, and J. Davis (2006). Spatial association between speciated fine particles and mortality. *Biometrics 62*(3), 855–863.

Gamerman, D. and H. F. Lopes (2006). *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. CRC Press.

Gelfand, A., L. Zhu, and B. Carlin (2001). On the change of support problem for spatio-temporal data. *Biostatistics 2*(1), 31.

Gelfand, A. E., S. Banerjee, and D. Gamerman (2005). Spatial process modelling for univariate and multivariate dynamic spatial data. *Environmetrics 16*(5), 465–479.

Gelfand, A. E., S. K. Sahu, and D. M. Holland (2012). On the effect of preferential sampling in spatial prediction. *Environmetrics 23*(7), 565–578.

Gelman, A. (2003). A Bayesian formulation of exploratory data analysis and goodness-of-fit testing. *International Statistical Review 71*(2), 369–382.

Gelman, A., J. Carlin, H. Stern, and D. Rubin (1995). *Bayesian Data Analysis*. Chapman and Hall, New York.

Gelman, A. and D. B. Rubin (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 457–472.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter (1996). Introducing Markov chain Monte Carlo. In *Markov chain Monte Carlo in practice*, pp. 1–19. Springer.

Gryparis, A., C. Paciorek, A. Zeka, J. Schwartz, and B. Coull (2009). Measurement error caused by spatial misalignment in environmental epidemiology. *Biostatistics 10*(2), 258–274.

Gu, C. (2013). *Smoothing spline ANOVA models*, Volume 297. Springer Science & Business Media, New York.

Gulliver, J., C. Morris, K. Lee, D. Vienneau, D. Briggs, and A. Hansell (2011). Land use regression modeling to estimate historic (1962- 1991) concentrations of black smoke and sulfur dioxide for Great Britain. *Environmental Science & Technology 45*(8), 3526–3532.

Guttorp, P. and P. Sampson (2010). Discussion of Geostatistical inference under preferential sampling by Diggle, P.J., Menezes, R. and Su, T. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 59*(2), 191–232.

Handcock, M. S. and J. R. Wallis (1994). An approach to statistical spatial-temporal modeling of meteorological fields. *Journal of the American Statistical Association 89*(426), 368–378.

Haneuse, S. J. et al. (2008). The combination of ecological and case–control data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(1), 73–93.

Haneuse, S. J. and J. C. Wakefield (2007). Hierarchical models for combining ecological and case–control data. *Biometrics 63*(1), 128–136.

Harvey, A. (1981). *Time series models*. Philip Allan. Oxford.

Hastie, T. J. and R. J. Tibshirani (1990). *Generalized Additive Models*. Number 43. CRC Press, Florida.

Illian, J. B., S. H. Sørbye, H. Rue, et al. (2012). A toolbox for fitting complex spatial point process models using integrated nested Laplace approximation (INLA). *The Annals of Applied Statistics 6*(4), 1499–1530.

Jekel, J. F., D. L. Katz, J. G. Elmore, and D. Wild (2007). *Epidemiology, biostatistics and preventive medicine*. Elsevier Health Sciences, Philadelphia, USA.

Jürgens, V., S. Ess, H. C. Phuleria, M. Früh, M. Schwenkglenks, H. Frick, T. Cerny, P. Vounatsou, et al. (2013). Bayesian spatio-temporal modelling of tobacco-related cancer mortality in Switzerland. *Geospatial Health 7*(2), 219–236.

Katsouyanni, K., G. Touloumi, E. Samoli, A. Gryparis, A. Le Tertre, Y. Monopolis, G. Rossi, D. Zmirou, F. Ballester, A. Boumghar, et al. (2001). Confounding and effect modification in the short-term effects of ambient particles on total mortality: results from 29 European cities within the APHEA2 project. *Epidemiology 12*(5), 521–531.

Kelsall, J. E., S. L. Zeger, and J. M. Samet (1999). Frequency domain log-linear models; air pollution and mortality. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 48*(3), 331–344.

Kleinschmidt, I., M. Hills, and P. Elliott (1995). Smoking behaviour can be predicted by neighbourhood deprivation measures. *Journal of Epidemiology and Community Health 49 (Suppl 2)*, S72–7.

Le, N. D., W. Sun, and J. V. Zidek (1997). Bayesian multivariate spatial interpolation with data missing by design. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 59*(2), 501–510.

Lee, B. L., M. R. Kosorok, and J. P. Fine (2005). The profile sampler. *Journal of the American Statistical Association 100*(471).

Lee, D., C. Ferguson, and E. M. Scott (2011). Constructing representative air quality indicators with measures of uncertainty. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 174*(1), 109–126.

Lee, D. and G. Shaddick (2010). Spatial modeling of air pollution in studies of its short-term health effects. *Biometrics 66*(4), 1238–1246.

Lee, J.-T., H. Kim, Y.-C. Hong, H.-J. Kwon, J. Schwartz, and D. C. Christiani (2000). Air pollution and daily mortality in seven major cities of Korea, 1991–1997. *Environmental Research 84*(3), 247–254.

Li, Y., P. Brown, H. Rue, M. al Maini, and P. Fortin (2012). Spatial modelling of lupus incidence over 40 years with changes in census areas. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 61*(1), 99–115.

Lindgren, F., H. Rue, and J. Lindström (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73*(4), 423–498.

Little, R. and D. Rubin (1987). *Statistical analysis with missing data*. John Wiley & Sons.

Logan, W. (1953). Mortality in the London fog incident, 1952. *The Lancet 261*(6755), 336–338.

Lumley, T. and L. Sheppard (2000). Assessing seasonal confounding and model selection bias in air pollution epidemiology using positive and negative control analyses. *Environmetrics 11*(6), 705–717.

Lunn, D., A. Thomas, N. Best, and D. Spiegelhalter (2000). WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing 10*(4), 325–337.

Mar, T. F., G. A. Norris, J. Q. Koenig, and T. V. Larson (2000). Associations between air pollution and mortality in Phoenix, 1995-1997. *Environmental Health Perspectives 108*(4), 347.

McCallum, E. and S. Weston (2011). *Parallel R*. O'Reilly Media, Inc.

McCullagh, P. and J. Nelder (1989). *Generalized Linear Models, 2nd Edition*. Chapman and Hall.

Moolgavkar, S. H. (2000). Air pollution and daily mortality in three US counties. *Environmental Health Perspectives 108*(8), 777.

Morris, C., J. Gulliver, D. Briggs, A. Hansell, et al. (2007). Modelling UK black smoke and sulphur dioxide concentrations, 1955-2001 to estimate lifecourse air pollution exposures. *Epidemiology 18*(5), S138.

Ott, W. (1990). A Physical Explanation of the Lognormality of Pollutant Concentrations. *Journal of the Air Waste Management Association 40*, 1378–1383.

Pati, D., B. J. Reich, and D. B. Dunson (2011). Bayesian geostatistical modelling with informative sampling locations. *Biometrika 98*(1), 35–48.

Peng, R. D. and M. L. Bell (2010). Spatial misalignment in time series studies of air pollution and health data. *Biostatistics 11*(4), 720–740.

Peters, A., J. Skorkovsky, F. Kotesovec, J. Brynda, C. Spix, H. E. Wichmann, and J. Heinrich (2000). Associations between mortality and air pollution in central Europe. *Environmental Health Perspectives 108*(4), 283.

Pope, C. A., D. W. Dockery, and J. Schwartz (1995). Review of epidemiological evidence of health effects of particulate air pollution. *Inhalation Toxicology 7*(1), 1–18.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys. New York*. Wiley & Sons.

Rue, H., S. Martino, and N. Chopin (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71*(2), 319–392.

Sahu, S. K., A. E. Gelfand, and D. M. Holland (2006). Spatio-temporal modeling of fine particulate matter. *Journal of Agricultural, Biological, and Environmental Statistics 11*(1), 61–86.

Sahu, S. K., A. E. Gelfand, and D. M. Holland (2007). High-resolution space–time ozone modeling for assessing trends. *Journal of the American Statistical Association 102*(480), 1221–1234.

Sahu, S. K. and K. V. Mardia (2005). A Bayesian kriged Kalman model for short-term forecasting of air pollution levels. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 54*(1), 223–244.

Samet, J. M., F. Dominici, F. C. Curriero, I. Coursac, and S. L. Zeger (2000). Fine particulate air pollution and mortality in 20 US cities, 1987–1994. *New England Journal of Medicine 343*(24), 1742–1749.

Schrödle, B. and L. Held (2011). A primer on disease mapping and ecological regression using INLA. *Computational Statistics 26*(2), 241–258.

Schwartz, J. (2000). The distributed lag between air pollution and daily deaths. *Epidemiology 11*, 320–326.

Schwartz, J. (2001). Air pollution and blood markers of cardiovascular risk. *Environmental Health Perspectives 109*(Suppl 3), 405.

Schwartz, J., D. Slater, T. V. Larson, W. E. Pierson, and J. Q. Koenig (1993). Particulate air pollution and hospital emergency room visits for asthma in Seattle. *American Review of Respiratory Disease 147*(4), 826–831.

Scott, A. and C. Wild (2001). Maximum likelihood for generalised case-control studies. *Journal of Statistical Planning and Inference 96*(1), 3–27.

Shaddick, G. and J. Wakefield (2002). Modelling daily multivariate pollutant data at multiple sites. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 51*(3), 351–372.

Shaddick, G. and J. V. Zidek (2014). A case study in preferential sampling: Long term monitoring of air pollution in the UK. *Spatial Statistics 9*, 51–65.

Smith, A. F. and G. O. Roberts (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, 3–23.

Spix, C., J. Heinrich, D. Dockery, J. Schwartz, G. Völksch, K. Schwinkowski, C. Cöllen, and H. E. Wichmann (1993). Air pollution and daily mortality in Erfurt, East Germany, 1980-1989. *Environmental health perspectives 101*(6), 518.

Stern, A. C. (1973). *Fundamentals of air pollution*. Elsevier, London.

Szpiro, A., L. Sheppard, and T. Lumley (2011). Efficient measurement error correction with spatially misaligned data. *Biostatistics 12*(4), 610–623.

Tonellato, S. F. (2001). A multivariate time series model for the analysis and prediction of carbon monoxide atmospheric concentrations. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 50*(2), 187–200.

Vanem, E., A. B. Huseby, and B. Natvig (2012). A Bayesian hierarchical spatio-temporal model for significant wave height in the North Atlantic. *Stochastic Environmental Research and Risk Assessment 26*(5), 609–632.

Vedal, S., M. Brauer, R. White, and J. Petkau (2003). Air pollution and daily mortality in a city with low levels of pollution. *Environmental Health Perspectives 111*(1), 45.

Verhoeff, A. P., G. Hoek, J. Schwartz, and J. H. van Wijnen (1996). Air pollution and daily mortality in Amsterdam. *Epidemiology 7*(3), 225–230.

Wahba, G. (1990). *Spline models for observational data*. Number 59. SIAM.

Wakefield, J. (2003). Sensitivity analyses for ecological regression. *Biometrics 59*(1), 9–17.

Wakefield, J. and R. Salway (2001). A statistical framework for ecological and aggregate studies. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 164*(1), 119–137.

Wakefield, J. and J.-P. H. Sebastien (2008). Overcoming ecologic bias using the two-phase study design. *American Journal of Epidemiology 167*(8), 908–916.

Wakefield, J. and G. Shaddick (2006). Health-exposure modeling and the ecological fallacy. *Biostatistics 7*(3), 438.

Waternaux, C., N. Laird, and J. Ware (1989). Methods for analysis of longitudinal data: blood lead concentrations and cognitive development. *Journal of the American Statistical Association 84*, 33–41.

Welty, L. J., R. Peng, S. Zeger, and F. Dominici (2009). Bayesian distributed lag models: estimating effects of particulate matter air pollution on daily mortality. *Biometrics 65*(1), 282–291.

WHO (1980). *The International Classification of Diseases, 9th Revision, Clinical Modification: ICD-9-CM*. Commission on Professional and Hospital Activities.

WHO (2011). Burden of disease attributable to outdoor air pollution. *World Health Organisation, Geneva*.

Wood, S. (2006). *Generalized additive models: an introduction with R*. CRC Press, London.

Zanobetti, A., M. Wand, J. Schwartz, and L. Ryan (2000). Generalized additive distributed lag models: quantifying mortality displacement. *Biostatistics 1*(3), 279–292.

Zhang, H. (2004). Inconsistent Estimation and Asymptotically Equal Interpolations in Model-Based Geostatistics. *Journal of the American Statistical Association 99*(465), 250–261.

Zhu, L., B. P. Carlin, and A. E. Gelfand (2003). Hierarchical regression with misaligned spatial data: relating ambient ozone and pediatric asthma ER visits in atlanta. *Environmetrics 14*(5), 537–557.

Zidek, J., L. Sun, N. Le, and H. Özkaynak (2002). Contending with space–time interaction in the spatial prediction of pollution: Vancouver's hourly ambient PM10 field. *Environmetrics 13*(5-6), 595–613.

Zidek, J., R. White, W. Sun, R. Burnett, and N. Le (1998). Imputing unmeasured explanatory variables in environmental epidemiology with application to health impact analysis of air pollution. *Environmental and Ecological Statistics 5*(2), 99–105.

Zidek, J. V., G. Shaddick, and C. G. Taylor (2014). Reducing estimation bias in adaptively changing monitoring networks with preferential site selection. *The Annals of Applied Statistics 8*(3), 1640–1670.